STATISTICAL ANALYSIS OF DIGITAL FIXED-POINT MULTIPLICATION ERRORS AND QUANTIZATION ERRORS

by

Leo P. Mulcahy

Sensor And Fire Control Department

December 1971

NAVAL UNDERSEA RESEARCH AND DEVELOPMENT CENTER, SAN DIEGO

NUC

Approved for public release; distribution unlimited.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)* Naval Undersea Research and Development Center San Diego, California 92132	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3 REPORT TITLE

Statistical Analysis of Digital Fixed-Point Multiplication Errors and Quantization Errors

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Research and Development; September 1969 – January 1970

5 AUTHOR(S) *(First name, middle initial, last name)*

Leo P. Mulcahy

6 REPORT DATE December 1971	7a. TOTAL NO. OF PAGES 55	7b. NO. OF REFS 12
8a. CONTRACT OR GRANT NO. SF 11-121-106, Task 8132 b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) NUC TP 254 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	

10 DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY Naval Ship Systems Command Washington, D. C. 20360

13 ABSTRACT

Deterministic properties of rounding and chopping were examined for both multiplication and quantization errors.

Statistical properties of rounding only were examined for both multiplication and quantization errors. Statistical properties examined are 1) the error distribution density (d.d.), 2) the error variance, 3) the autocorrelation between successive error values, and 4) the cross-correlation coefficient between the quantizer input and the resulting error. Specific results were obtained for zero-mean Gaussian random processes.

For quantization errors the above properties depend only on the ratio of the process standard deviation to the quantization interval size (σ/q). The mapping of the quantizer input d.d. onto the quantization error d.d. is continuous. Consequently, for $\sigma/q \geqslant 1.0$, the error d.d. is almost exactly uniform between $\pm q/2$, and the error variance is very near $q^2/12$. Both the autocorrelation and the cross-correlation coefficients weie negligible. Furthermore, the equations show that the quantization error approaches arbitrarily close to $q^2/12$ as σ/q increases, while the autocorrelation and cross-correlation coefficients approach arbitrarily close to zero.

For rounding errors the above properties depend not only on σ/q, but on the word size, N, and the value of the multiplier, J, as well. Furthermore, the discrete nature of the computer word causes a discrete mapping of the multiplier input d.d. onto the multiplication error d.d. Consequently, for the limited range of parameters considered, most values of J yield an error d.d. which is not uniform in the continuous sense but shows a variance approaching $q^2/12$. Similarly, most autocorrelation and cross-correlation values approach zero, but stabilize at some non-zero value as σ/q becomes large. However, some values of J result in large, non-zero autocorerelation and cross-correlation values and a variance which diverges widely from $q^2/12$.

NAVAL UNDERSEA RESEARCH AND DEVELOPMENT CENTER, SAN DIEGO, CA. 92132

AN ACTIVITY OF THE NAVAL MATERIAL COMMAND

CHARLES B. BISHOP, Capt., USN
Commander

Wm. B McLEAN, Ph.D.
Technical Director

The work reported was done from September 1969 to January 1970 under SF11-121-106, Task 8132.

Released by
D. A. HANNA, Acting Head
Simulation Analysis and
Applications Division

Under authority of
W. D. SQUIRE, Head
Sensor and Fire Control
Department

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
statistical theory						
error analysis						
autocorrelation						
cross-correlation						
variance						
analog to digital converters						

SUMMARY

PROBLEM

Review and extend the theory of computation error generation in digital filters. Specifically consider fixed-point multiplication for digitized Gaussian analog signal inputs.

RESULTS

Deterministic properties of rounding and chopping were examined for both multiplication and quantization errors.

Statistical properties of rounding only were examined for both multiplication and quantization errors. Statistical properties examined are 1) the error distribution density (d.d.), 2) the error variance, 3) the autocorrelation between successive error values, and 4) the cross-correlation coefficient between the quantizer input and the resulting error. Specific results were obtained for zero-mean Gaussian random processes.

For quantization errors, the above properties depend only on the ratio of the process standard deviation to the quantization interval size (σ/q). The mapping of the quantizer input d.d. onto the quantization error d.d. is continuous. Consequently, for $\sigma/q > 1.0$, the error d.d. is almost exactly uniform between $\pm q/2$, and the error variance is very near $q^2/12$. Both the autocorrelation and the cross-correlation coefficients were negligible. Furthermore, the equations show that the quantization error approaches arbitrarily close to $q^2/12$ as σ/q increases, while the autocorrelation and cross-correlation coefficients approach arbitrarily close to zero.

For rounding errors, the above properties depend not only on σ/q, but on the word size, N, and the value of the multiplier, J, as well. Furthermore, the discrete nature of the computer word causes a discrete mapping of the multiplier input d.d. onto the multiplication error d.d. Consequently, for the limited range of parameters considered, most values of J yield an error d.d. which is not uniform in the continuous sense but shows a variance approaching $q^2/12$. Similarly, most autocorrelation and cross-correlation values approach zero, but stabilize at some non-zero value as σ/q becomes large. However, some values of J result in large, non-zero autocorrelation and cross-correlation values and a variance which diverges widely from $q^2/12$.

RECOMMENDATIONS

1. Attempt to derive analytical formulas for evaluation of the error variance and the cross-correlation between the multiplier input and the resulting error.

2. Extend the technique of analysis to the problem of error generation for floating-point computers.

CONTENTS

INTRODUCTION

OBJECTIVE

We consider errors that occur when two fixed-point binary numbers are multiplied in a digital computer. For example, suppose we multiply the following numbers together, 0.0101_2 and 1.1001_2*. Each number has a "word" length of N = 4 bits plus sign bit.

The result of multiplication is a number 2 × 4 = 8 bits plus sign bit, equal to 1.00101101_2 in this case. To store the result in the computer memory it is necessary to "chop" or "round" it by discarding the 4 least significant bits. Chopping results when we discard these bits, leaving the first 4 bits plus sign unchanged. Rounding is done by adding a "one" to the least significant of the first 4 bits when the most significant of the lower 4 bits is equal to a "one". Then the lower 4 bits are discarded. If the most significant of the lower 4 bits is equal to a "zero" nothing is added before the lower 4 bits are discarded. In the above example, chopping would leave the result 1.0010_2; rounding would leave the result 1.0011_2. This example and another are shown in detail in table 1.

Note that chopping or rounding results in a product that is inexact. The difference between the chopped or rounded product and the original product is an error. It is with this error and its characteristics that we are concerned.

Multiplication errors must be considered in evaluating the performance of digital filters. A digital filter is an algorithm which is used in a digital computer to replace an analog filter (reference 2). The algorithm is of the form

$$y(nT) = \sum_{i=1}^{n} K_i y(nT - iT) - \sum_{i=0}^{m} L_i x(nT - iT), \qquad n = 0, 1, 2, \ldots \qquad (1)$$

*The subscript refers to the base of the number system used. In this case the base is 2, or the binary number system. The first number is $0.0101_2 = (-1)^0 \cdot (0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4}) = +0.3125_{10}$. The second number is $1.1001_2 = (-1)^1 \cdot (1 \times 2^{-1} + 0 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4}) = -0.5625_{10}$. The period is called a binary point in the binary number system, and a decimal point in the decimal number system (reference 1).

Table 1. Examples of Rounding and Chopping.

	EXAMPLE NO. 1	EXAMPLE NO. 2
NUMBER A	0.0101	0.0101
NUMBER B	1.1001	0.1010
A TIMES B	1.00101101	0.00110010
RESULTS AFTER		
ROUNDING	1.0011	0.0011
CHOPPING	1.0010	0.0011
VALUE OF ERROR		
ROUNDING	1.00000011	1.00000010
CHOPPING	0.00001101	1.00000010

FROM EXAMPLE NO. 1

A TIMES B = 1.00101101

FOR ROUNDING: A "ONE" IS PRESENT IN THIS BIT POSITION. SO WE ADD A "ONE" TO THE NEXT HIGHER BIT POSITION AND DISCARD THE LOWER 4 BITS. THE RESULT IS EQUAL TO 1.0011. THE SIGN OF THE 8-BIT WORD IS IGNORED IN THIS OPERATION.

FOR CHOPPING: THE LOWER 4 BITS ARE DISCARDED.

NOTE: ALL NUMBERS IN THIS EXAMPLE ARE BINARY. THE ERROR IS DEFINED AS THE QUANTITY THAT IS ADDED TO (A TIMES B) TO GET THE ROUNDED OR CHOPPED RESULT.

where $x(nT)$, $n = 0, 1, 2, \ldots$, is a sequence of numbers obtained from the analog waveform, $\tilde{x}(t)$. This number sequence is the input to the digital filter. The resulting output sequence from the digital filter is $y(nT)$, $n = 0, 1, 2, \ldots$. The input sequence, $x(nT)$, results from periodic sampling of $\tilde{x}(t)$ at a rate $f_s = 1/T$, and subsequent conversion of these samples to digital numbers (analog-to-digital conversion). Figure 1 is a block diagram of the operations needed to produce a digital filter which is equivalent to the analog filter.

Equation (1) requires three arithmetic operations: multiplication, addition, and subtraction. Chopping or rounding errors occur only in multiplication. Overflow errors occur in addition and subtraction.* Note also that the coefficients K_i and L_i canno' be specified exactly, in general, due to the finite word length in the computer. Consequently, the desired (unquantized) coefficient value differs by a fixed amount from the quantized coefficient value, and the filter does not have the characteristics desired. This may be a problem if the filter is sensitive to small changes in the value of the coefficient. Furthermore, the recursive equation leads to dead-band and other effects (reference 2). We will not discuss these effects further since we are only concerned with the problem of multiplication error generation.

This report consists of three major sections. The first section, Quantization Errors, reviews the theory of quantization errors, which forms the foundation of the theory of multiplication errors used until now. The second section, Multiplication Errors, defines the deterministic and statistical properties of multiplication errors and compares multiplication errors with quantization errors. The final section, Summary, summarizes the analysis performed in the second section and suggests an approach to further analysis.

BACKGROUND

Multiplication errors are similar to errors that occur when an analog waveform sample is quantized. It seems natural, therefore, to extend the conclusions of the analysis of quantization errors to the analysis of multiplication errors. Quantization errors have been analyzed by Bennett (reference 3), Widrow (reference 4), and Shaver (reference 5). Bennett found that quantization is equivalent to adding an independent (white) noise that is uniformly distributed over the quantization interval q, to the original (unquantized) samples.** He also determined the autocorrelation of successive samples of the quantization errors as a function of the autocorrelation of successive samples of the analog waveform. The analog waveform was Gaussian. Widrow determined the difference between the actual variance of the quantization error and the variance of the uniformly distributed

*An overflow occurs when the sum of two numbers is too large for the computer word length. For example, the sum of 0.1010_2 and 0.1000_2 is not representable. The carry generated by the sum has no place to go since the bit position in front of the binary point is used to denote the sign of the number.

**q is the size of the basic quantization interval. Complete definition of the quantization process is given later in this paper.

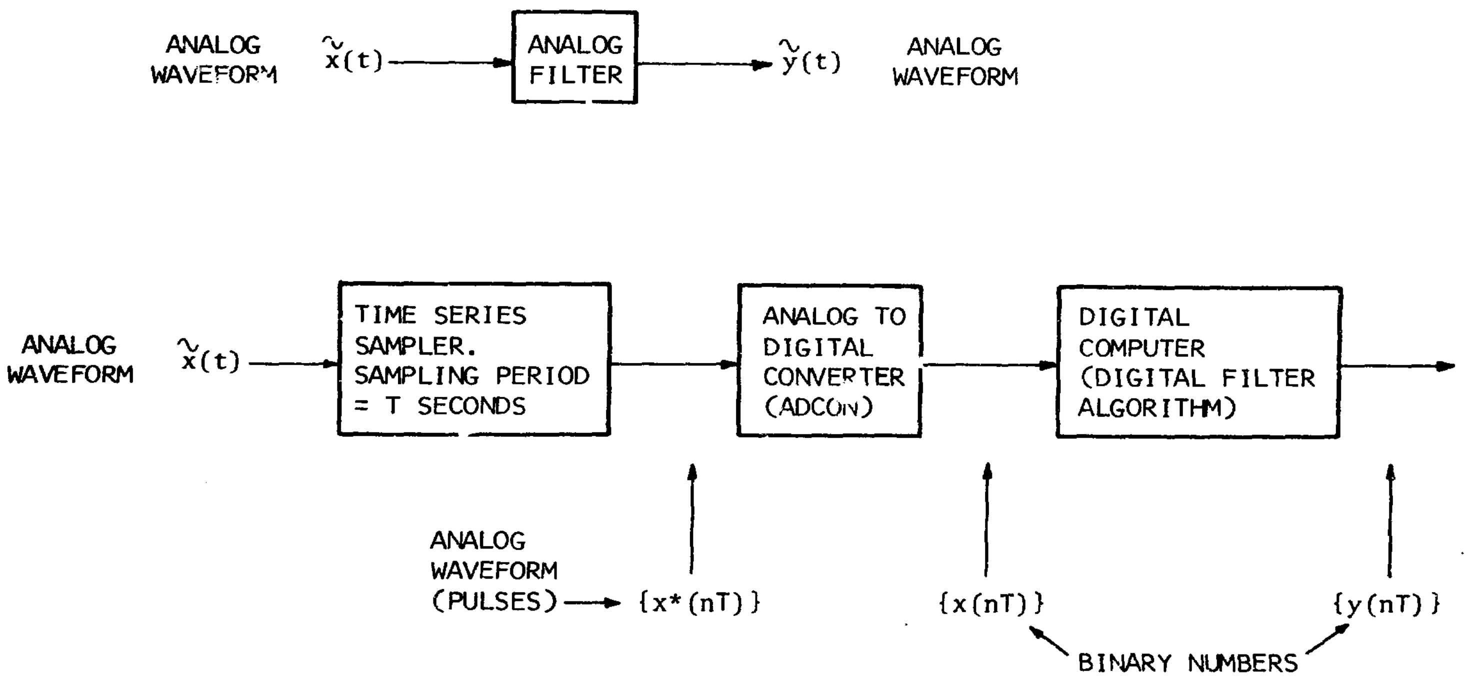

Figure 1. Block diagrams of an analog filter and an equivalent digital filter.

quantization error. Shaver later derived the cross-correlation between the quantizer input and quantization error at the same instant. Both analyses were for a Gaussian analog waveform.

As we shall see, chopping and rounding are exactly the same operation for multiplication errors as for quantization errors. This has been noted in work on the effects of multiplication errors in digital computation. For example, Knowles and Edwards (references 6, 7, 8) assumed that multiplication errors could be treated as an additive, independent white noise because they are similar to quantization errors. They analyzed the effects of this noise on the steady-state performance of digital control systems of the form given in Equation (1). As a chec on their assumptions, they computed the autocorrelation functions of some multiplication error sequences (reference 6, p. 2384). A number of word lengths, multiplication coefficients, and sampled analog waveforms were used. They concluded that multiplication roundoff error spectra are essentially white with respect to practical sampled-data systems. Unfortunately, the paper was not clear as to what kind of sampled-data system was used for the measurements. Also, the rms value of the analog signal relative to the quantization interval was not given.

Gold and Rader (reference 9) also linked quantization errors with multiplication errors, referring to the work of Bennett. They experimentally verified the mean-square output noise for a one-pole digital filter as a function of the pole position and word length (28 and 29 bits including sign). They did not compute correlation of successive output errors, nor give the rms value of the analog signal. Gold and Rabiner (reference 10) essentially followed Gold and Rader in their assumptions.

All investigators experimentally verified the assumed similarity of quantization errors to multiplication errors by measuring the mean square output noise of a finite wordlength digital filter. Only Knowles and Edwards (reference 6) computed correlation of error sequences, and did so for specific systems and specific word sizes only.

QUANTIZATION ERRORS

Later in this report we will compare the statistical properties of quantization errors with those of multiplication errors. This section is a review of theoretical results needed for such a comparison.

QUANTIZER CHARACTERISTICS: DETERMINISTIC

A quantizer is used in an analog-to-digital converter (ADCON). An ADCON converts time series samples of an analog signal to digital form so they can be accepted by a digital computer. A block diagram of the process is shown in figure 2. The block diagram does not represent the circuit operations, but shows the equivalent operations in a mathematical sense. We will discuss quantization first and then cover the other necessary aspects of the ADCON.

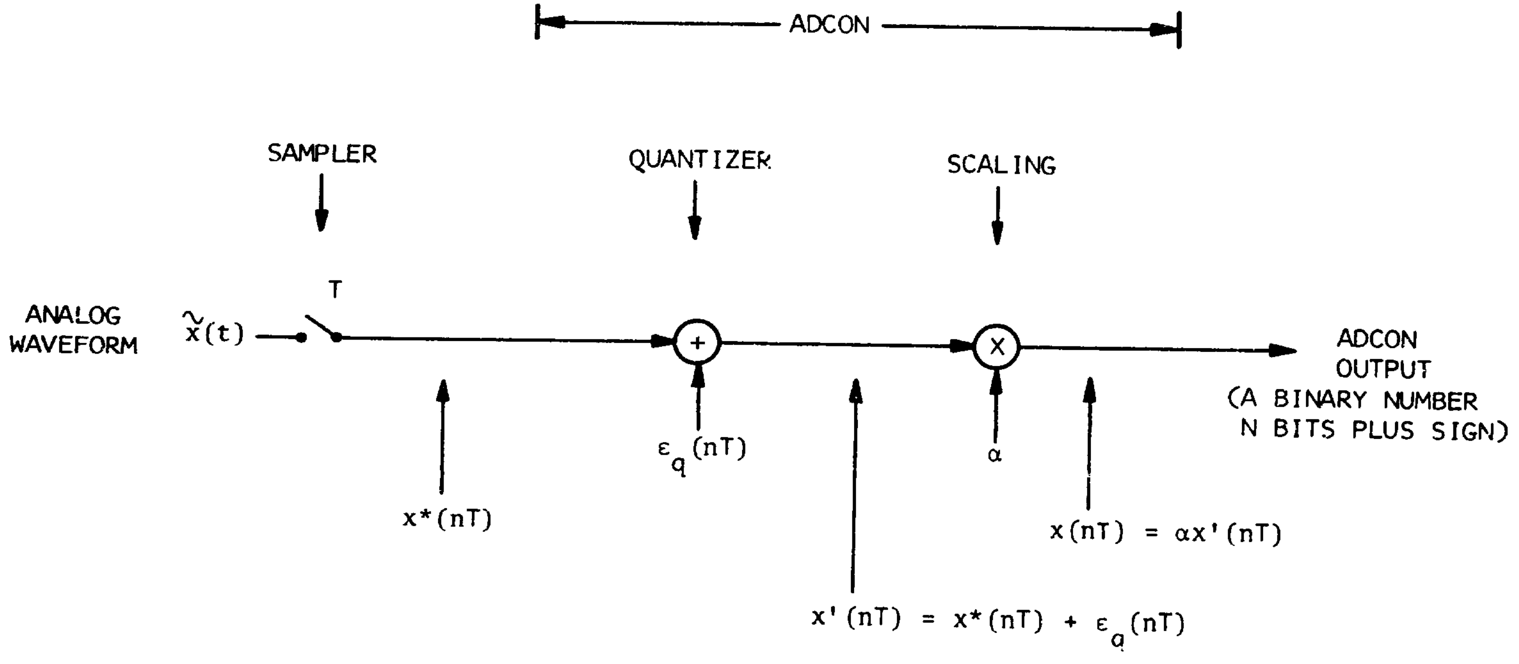

Figure 2. Block diagram representation of a sampler and analog-to-digital converter.

In figure 2 the sampler output voltage $x^*(nT)$ is a random variable whose probability density function is continuous in the interval $-B < x^*(nT) \leqslant A$. $-B$ and A are lower and upper limits on the range of $x^*(nT)$ which are set by the physical nature of the sampler and quantizer.

Quantization is a process of subdivision of the range of $x^*(nT)$ into class intervals. For rounding,

$$(i - \frac{1}{2})q < x^*(nT) \leqslant (i + \frac{1}{2})q, \tag{2}$$

where

$$i = -\frac{B}{q}, -\frac{B}{q} + 1, \ldots, -1, 0, 1, \ldots, \frac{A}{q} - 1, \frac{A}{q}$$

(A and B are chosen to be integer multiples of q).

Each class interval is of equal width q. The quantizer output voltage is iq. This is the center value of the class interval. For example, if $(3/2)q < x^*(nT) \leqslant (5/2)q$, the quantizer output is 2q volts.

The difference between the quantizer output $x'(nT)$ and the input $x^*(nT)$ is the quantization error $\epsilon_q(nT)$.

That is,

$$\epsilon_q(nT) = x'(nT) - x^*(nT). \tag{3}$$

(Sometimes the error is defined as the input minus the output. Our convention assumes that the quantization error is added to the quantizer input.) In the case of equation (2) this represents a rounding error since the input samples $x^*(nT)$ are rounded to the nearest class interval center value. We will call the sequence of errors $\epsilon_q(nT)$, $n = 0, 1, 2, \ldots$, quantization noise or the quantization process. Note that $|\epsilon_q(nT)| \leqslant q/2$.

For chopping,

$$iq \leqslant |x^*(nT)| < (i + 1)q, \tag{4}$$

where

$$i = 0, 1, \ldots, A/q.$$

The quantizer output voltage is (sgnx*)iq, where

$$\mathrm{sgn}x^* = \begin{cases} 1, & x^* > 0 \\ -1, & x^* < 0, \end{cases} \tag{5}$$

and iq is the value of the lower end of the class interval defined in equation (4). The sign of the quantizer output voltage is the same as the sign of $x^*(nT)$. For example, if $3q \leqslant |x^*(nT)| < 4q$ and $x^*(nT) < 0$, the quantizer output is $-3q$ volts. Note that $|\epsilon_q(nT)| \leqslant q$.

Quantizer input/output characteristics for both rounding and chopping are shown in figure 3. The equivalent ADCON outputs are also shown. The quantization error is a deterministic function of the quantizer input. The functions can be written as follows:

Rounding:

$$\epsilon_q(nT)|_{x^*} = -x^* + iq, \tag{6}$$

where

$$i = 0, \pm 1, \pm 2, \ldots$$

and

$$-q/2 \leqslant \epsilon_q(nT)|_{x^*} < q/2.$$

Chopping:

$$\epsilon_q(nT)|_{x^*} = -x^* + iq \quad \begin{cases} i = 0, 1, 2, \ldots, \quad x^* \geqslant 0 \\ -q < \epsilon_q(nT)|_{x^*} \leqslant 0 \\ i = 0, -1, -2, \ldots, \quad x^* \leqslant 0 \\ 0 \leqslant \epsilon_q(nT)|_{x^*} < q. \end{cases} \tag{7}$$

These functions are shown in figure 4.

Figure 5 is an example of an analog signal and the results of sampling and quantization. Results for both rounding and chopping are shown. A couple of features are

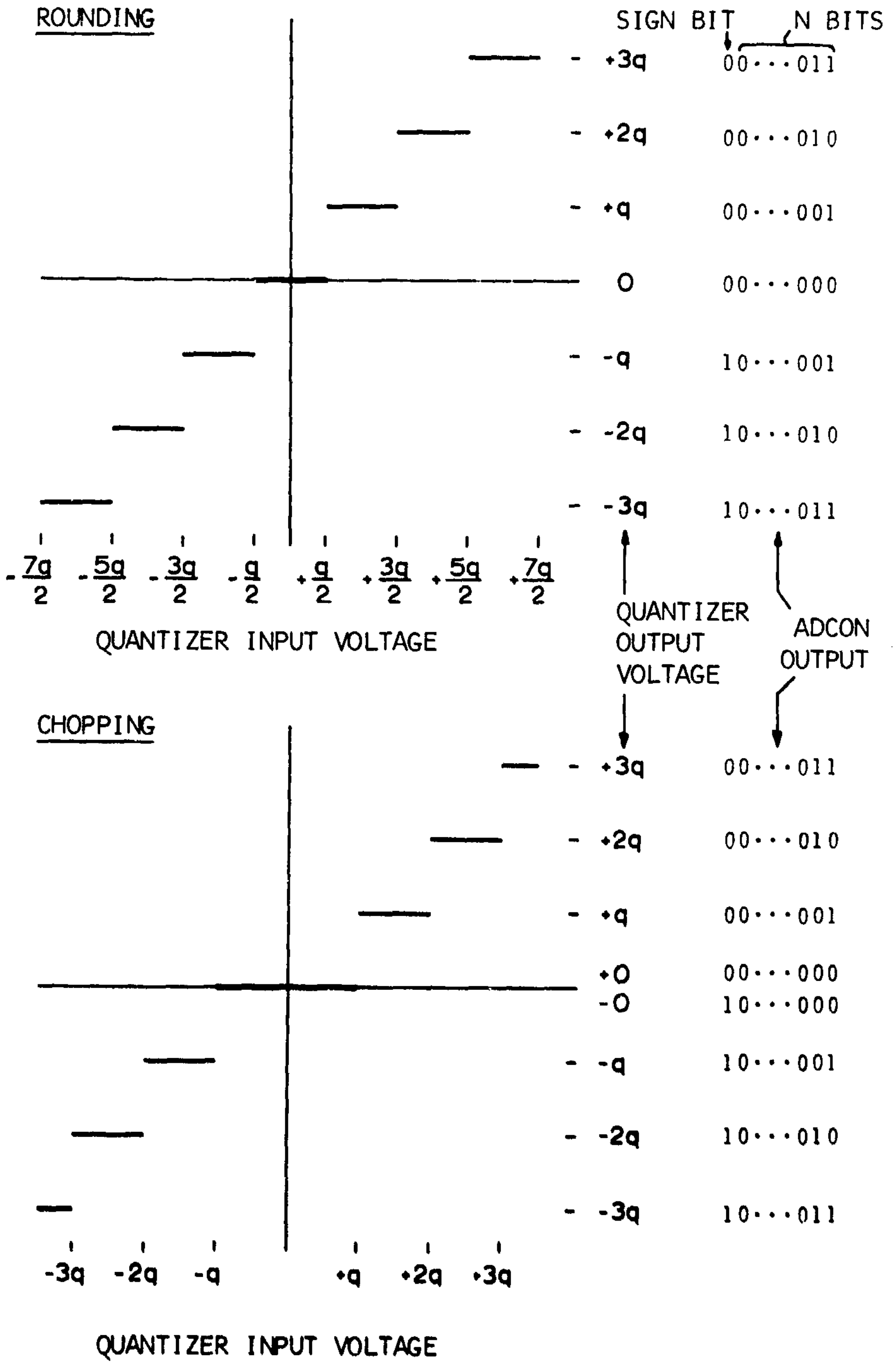

Figure 3. Quantizer input/output characteristics.

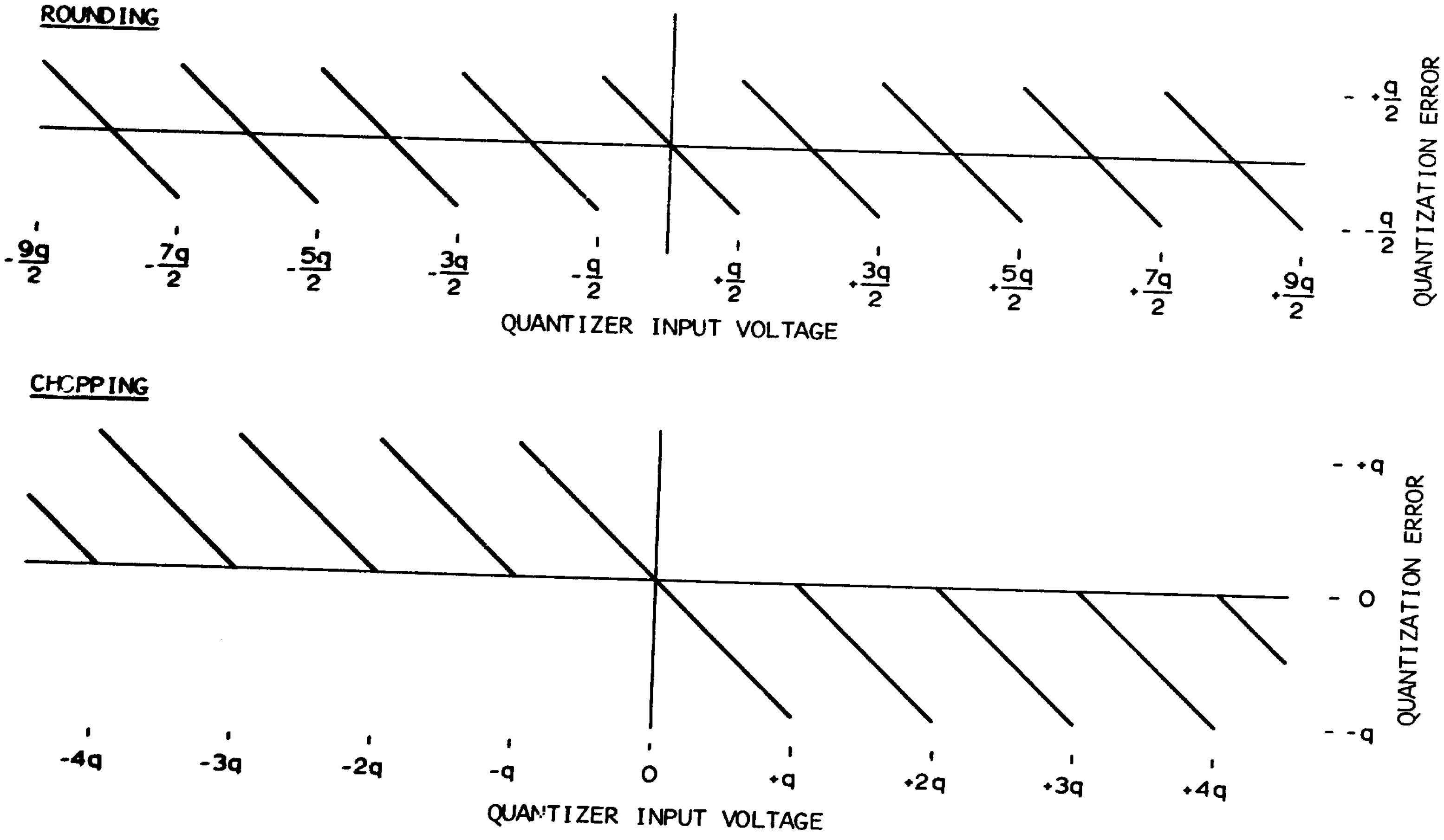

Figure 4. Quantization error as a function of the quantizer input voltage.

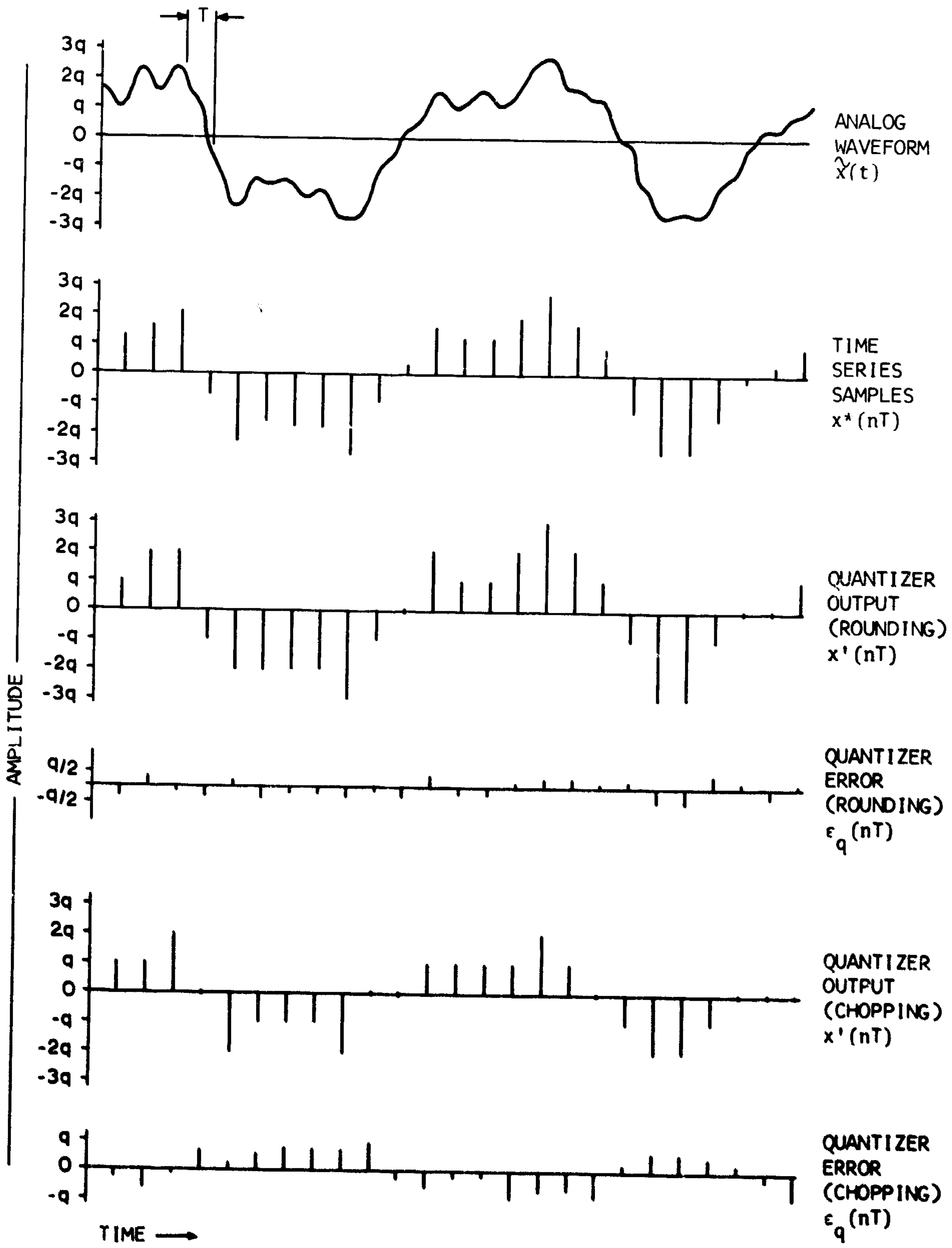

Figure 5. Example of an analog waveform and quantizer outputs showing error sequences.

evident. First, the chopping errors are generally larger than the rounding errors. The rounding-error sequence is bounded by ±q/2 and the chopping error sequence bounded by ±q. Second, the chopping-error sequence is quasi-periodic and dependent on the polarity of the analog signal input. The quasi periodicity can be eliminated by adding a d.c. level to the analog signal. This results in a d.c. bias in the error sequence, though.

Analog-to-digital conversion results in a binary number, not an analog voltage. The placement of a binary point in this number is not relevant as far as ADCON operation is concerned. It becomes important to the user because scaling is necessary in order to interpret the ADCON number properly. For our purpose we will assume a leading binary point. Then, if the quantizer output voltage is $x'(nT) = iq$, where $i = \ldots -1, 0, +1, \ldots$, the corresponding ADCON output is just $x(nT) = i2^{-N}$. There is a multiplicative relation α between $x'(nT)$ and $x(nT)$. That is, $x(nT) = \alpha x'(nT)$. If $x'(nT) = q$, then $x(nT) = 2^{-N}$. Thus, $\alpha = 2^{-N}q^{-1}$. This scale factor is included in figure 2. Note that the size of the least significant bit at the ADCON output is equivalent to q when referred back to the ADCON input.

QUANTIZER CHARACTERISTICS: STATISTICAL

We will consider the following questions about the statistical characteristics of quantization errors and quantization error sequences:

1. What shape does the quantization error distribution density (d.d.) have?
2. What are the mean and variance of quantization errors?
3. What conditions must apply for the error sequence to be considered a source of white noise?
4. What conditions must apply for the error sequence to be uncorrelated with the signal sequence?

Distribution Densities of Quantization Errors

We will show how the quantizer input d.d. determines the quantization error d.d. For rounding, the probability that the quantizer output is equal to iq is

$$\begin{aligned} P_{x'}(iq) &= \text{Prob}\left\{x' = iq\right\} \\ &= \text{Prob}\left\{\left(i - \tfrac{1}{2}\right)q < x^* \leq \left(i + \tfrac{1}{2}\right)q\right\} \\ &= \int_{(i-1/2)q}^{(i+1/2)q} p_{x^*}(x^*)\,dx^* \qquad (8) \end{aligned}$$

where $p_{x^*}(x^*)$ is the d.d. of the quantizer input. The argument of the random variable x^* (the sequence index nT) is omitted since the sequence is assumed stationary.

The joint d.d. of the quantizer input x^* and the quantizer error ϵ_q is

$$p_{\epsilon_q\, x^*}(\epsilon_q, x^*) = p_{\epsilon_q | x^*}(\epsilon_q | x^*) p_{x^*}(x^*)$$

$$= \sum_{i=-\infty}^{\infty} \delta(\epsilon_q - [iq - x^*]) p_{x^*}(x^*), \qquad (9)$$

where $-q/2 \leq \epsilon_q < q/2$ and $\delta(x)$ is a delta function. We define $\delta(x)$ as a distribution which assigns to a continuous function $\phi(t)$ the number $\phi(o)$. That is, we use the special integral definition (reference 11, pp. 269-282)

$$\int_{-\infty}^{\infty} \delta(t)\phi(t)\, dt = \phi(o).$$

We assign infinite limits to the summation for the sake of simplicity. In reality the limits are given by equation (2).

The d.d. of the quantizer error is then

$$p_{\epsilon_q}(\epsilon_q) = \int_{-\infty}^{\infty} p_{\epsilon_q\, x^*}(\epsilon_q, x^*)\, dx^*$$

$$= \int_{-\infty}^{\infty} \left\{ \sum_{i=-\infty}^{\infty} \delta(\epsilon_q - [iq - x^*]) p_{x^*}(x^*) \right\} dx^*$$

$$= \sum_{i=-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \delta(\epsilon_q - [iq - x^*]) p_{x^*}(x^*) \right\} dx^*$$

$$= \sum_{i=-\infty}^{\infty} p_{x^*}(iq - \epsilon_q), \qquad -q/2 \leq \epsilon_q < q/2. \qquad (10)$$

For chopping, the probability that the quantizer output is equal to iq is

$$P_{x'}(iq) = \begin{cases} \displaystyle\int_{iq}^{(i+1)q} p_{x^*}(x^*)\, dx^*, & \begin{array}{l} x^* \geqslant 0 \\ i = 0, 1, \ldots \end{array} \\ \displaystyle\int_{(i-1)q}^{iq} p_{x^*}(x^*)\, dx^*, & \begin{array}{l} x^* < 0 \\ i = 0, -1, \ldots \end{array} \end{cases} \tag{11}$$

The joint d.d. of x^* and ϵ_q is

$$p_{\epsilon_q\, x^*}(\epsilon_q, x^*) = \sum_{\substack{i=0 \\ x^*>0 \\ -q<\epsilon_q\leqslant 0}}^{\infty} \delta(\epsilon_q - [iq - x^*]) p_{x^*}(x^*) + \sum_{\substack{i=0 \\ x^*<0 \\ 0\leqslant\epsilon_q<q}}^{-\infty} \delta(\epsilon_q - [iq - x^*]) p_{x^*}(x^*). \tag{12}$$

(See equation 7)

The same procedure yields the following d.d. of the quantization error.

$$p_{\epsilon_q}(\epsilon_q) = \sum_{\substack{i=0 \\ -q<\epsilon_q<0}}^{\infty} p_{x^*}(iq - \epsilon_q) + \sum_{\substack{i=0 \\ 0<\epsilon_q<q}}^{-\infty} p_{x^*}(iq - \epsilon_q). \tag{13}$$

(We ignore the terms for $\epsilon_q = 0$. They have the effect of introducing a delta function in the d.d. for ϵ_q. This is because of the limits on ϵ_q imposed by equations (2) and (4). We could have chosen the limits so that the d.d. for ϵ_q is continuous at $\epsilon_q = 0$, but this would have further complicated the presentation.)

Equations (10) and (13) show that the quantization error d.d. results from a mapping of the quantizer input d.d.. The mapping equations are (6) and (7). The nature of the mapping is illustrated in figures 6 and 7. Each figure shows the following d.d.'s: quantizer input, quantizer output, quantization error, and the joint d.d. of the quantizer input and the quantization error. A zero-mean Gaussian d.d. is shown for the quantizer input. $\sigma/q = 1$, where σ^2 is the variance of the d.d..

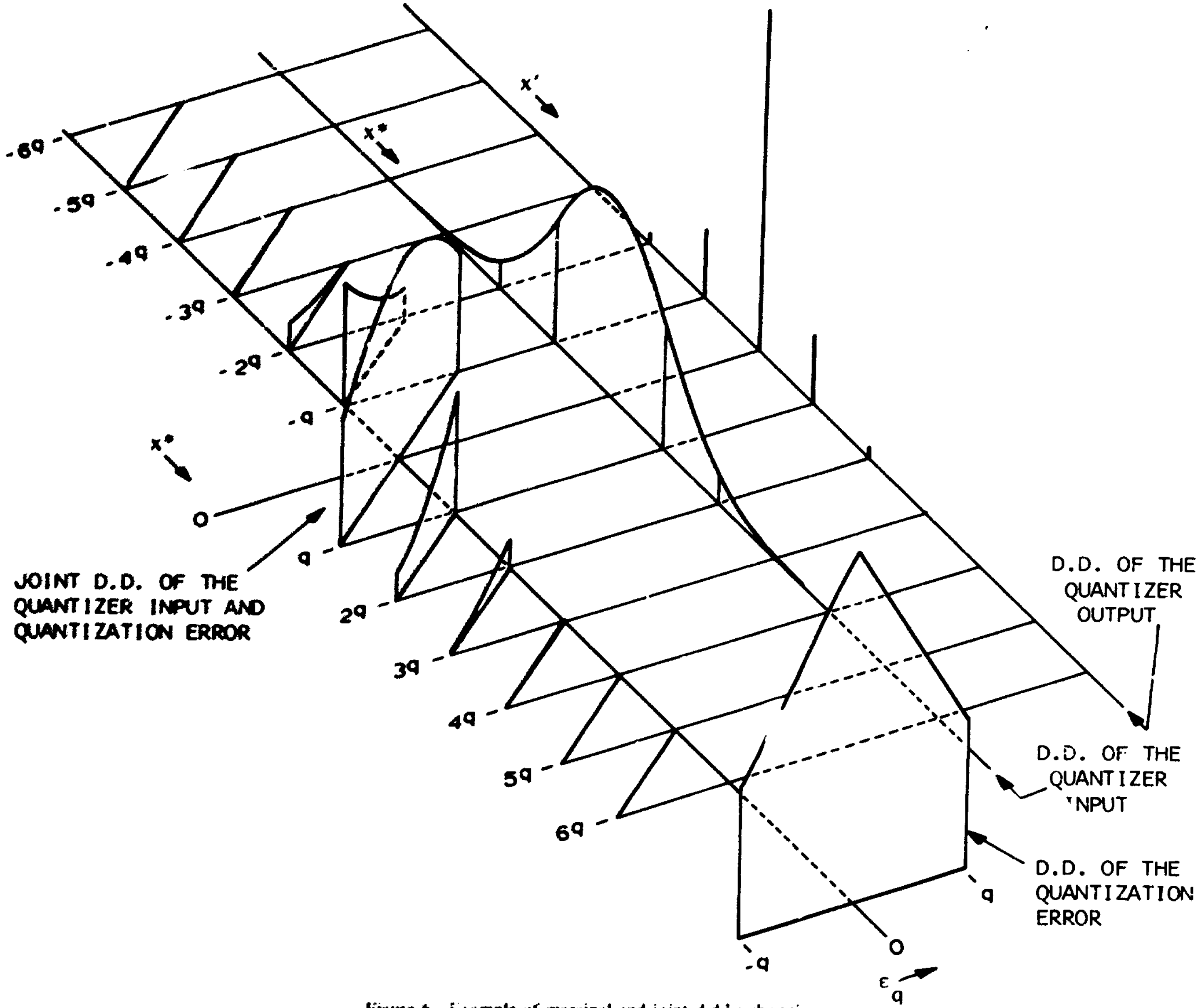
-6q
-5q
-4q
-3q
-2q
-q
x*
O
q
2q
3q
4q
5q
6q
x*
x'
JOINT D.D. OF THE QUANTIZER INPUT AND QUANTIZATION ERROR
D.D. OF THE QUANTIZER OUTPUT
D.D. OF THE QUANTIZER 'NPUT
D.D. OF THE QUANTIZATION ERROR
-q
O
q
ε_q

Figure 6. Example of marginal and joint d.d.'s; chopping.

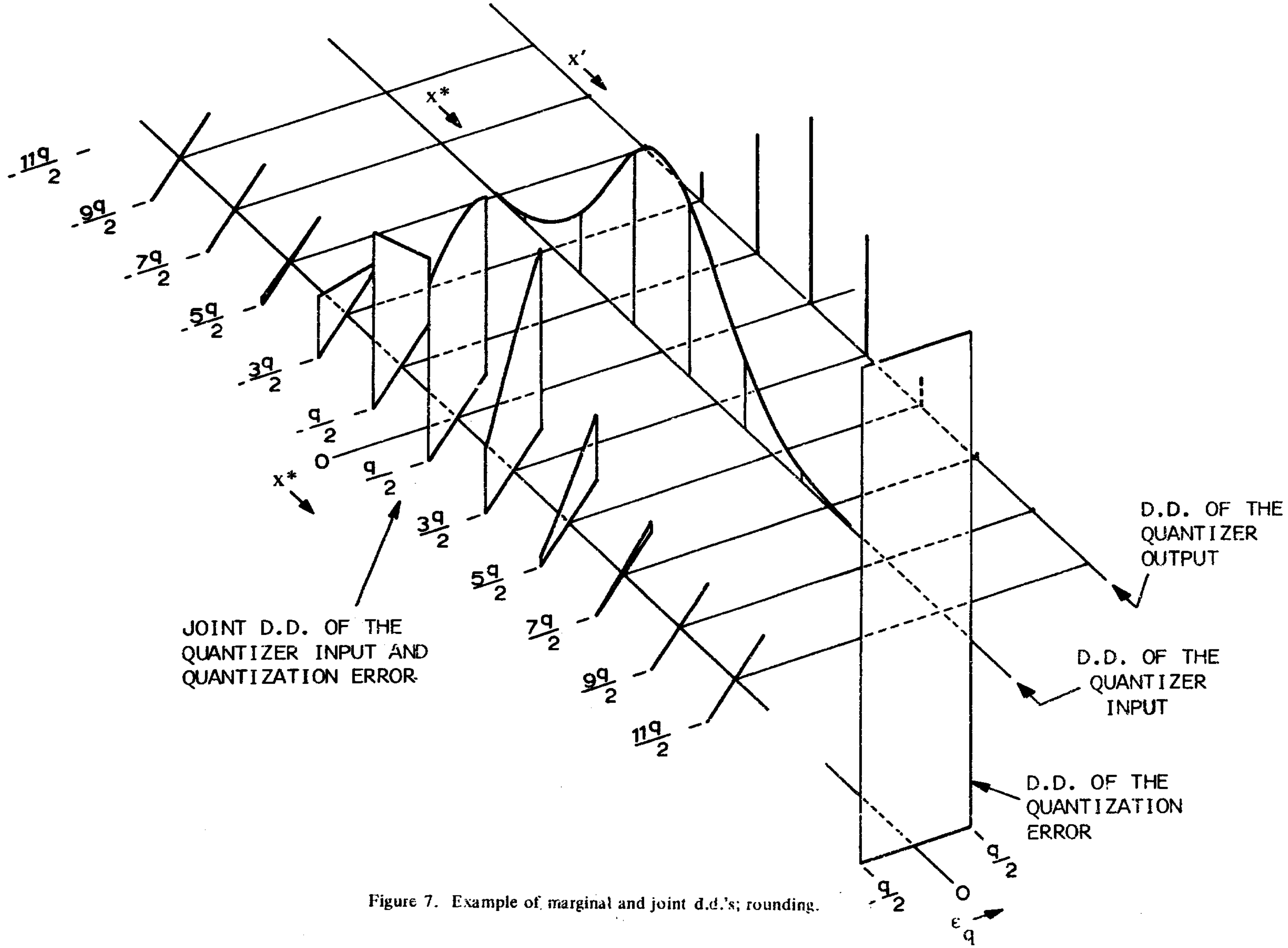

Figure 7. Example of marginal and joint d.d.'s; rounding.

The error d.d. for chopping is wider than that for rounding. Also, the error d.d. is definitely not uniform for chopping, but is for rounding. The d.d. for rounding is actually not uniform, but is so nearly uniform that the graph scale won't show the variations. The d.d. for chopping will become more uniform as σ/q is increased. However, the quantization error will always be negatively correlated with the quantizer input. We consider only rounding from now on, except when we look at multiplication errors.

Widrow's Results

The following quotation is from a summary in reference 12 by Widrow (see also reference 4).

> *The probability density of a quantizer output signal is discrete, consisting of a series of uniformly separated impulses with spacing equal to the quantization box size q. This density has a characteristic function (Fourier transform) which is periodic with a "frequency" $\phi = 2\pi/q$. A comparison of quantization with the addition of an independent uniformly-distributed (between ±q/2) noise shows that the quantizer output distribution density consists of samples of the distribution density of signal plus noise. Satisfaction of a quantizing theorem ensures that statistics can be recovered from quantized samples and that quantization noise itself is precisely flat-topped distributed. The quantization of high-order (correlated) signals compares with the addition of first-order noise (statistically independent, white). When a multidimensional quantizing theorem is satisfied, quantization noise is first-order and uncorrelated even though signals may be highly correlated.*

The quantizing theorem referred to says essentially the following: suppose the "frequency" $\phi = 2\pi/q$ is twice as high as the "highest frequency component" contained in the shape of the quantizer input d.d. $p_x*(x*)$. It is then possible to recover $p_x*(x*)$ from the quantizer output d.d. $p_{x'}(iq)$.

In most cases the theorem is not completely satisfied. The quantization error is then only approximately uniformly distributed between ±q/2. The quantization error variance will be in error by some amount. Widrow obtained a formula for the error of the quantization error. This was for a zero-mean Gaussian quantizer input. Rounding only was treated. The error is, to a close approximation,

$$\epsilon_\nu = 2\sigma^2 \exp(-2\pi^2\sigma^2/q^2)\,(2 + \frac{q^2}{\sigma^2}\ \frac{1}{2\pi^2}) \qquad (14)$$

The proportion error is

$$\epsilon'_\nu = \frac{\epsilon_\nu}{q^2/12}$$

$$= 24\frac{\sigma^2}{q^2} \exp\left(-2\pi^2\sigma^2/q^2\right)\left(2 + \frac{q^2}{\sigma^2} - \frac{1}{2\pi^2}\right) \tag{15}$$

This equation is a monotonic decreasing function of σ/q. For example, the proportion error for $\sigma/q = 1.0$ is $\epsilon'_\nu < 1.5 \times 10^{-7}$, a very small number. This means that the quantization error d.d. is almost uniform for a low value of σ/q.

Bennett's Results

Bennett (reference 3) discussed the distortion effect of sampling and quantization on analog waveforms. One of his results is the following (reference 3, p. 455).

> *Distortion caused by quantizing errors produces much the same sort of effects as an independent source of noise. The reason for this is that the spectrum of the distortion in the receiving filter output is practically independent of that of the signal over a wide range of signal magnitudes. Even when the signal is weak so that only a few quantizing steps are operated there is usually enough residual noise on actual systems to determine the quantizing noise and mask the relation between it and the signal.*

Bennett obtained a lengthy formula for the autocorrelation of the quantization errors as a function of the autocorrelation of the analog waveform. This was for a zero-mean Gaussian quantizer input. He reduced the formula to an accurate approximation which we reproduce here (reference 3, p. 467, equation (2.26)). The notation is changed to conform to this report.

$$\frac{R_{\epsilon_q\epsilon_q}(\tau)}{R_{x*x*}(0)} = \frac{q^2}{\sigma^2}\,\frac{1}{2\pi^2}\sum_{m=1}^{\infty}\frac{1}{m^2}\exp\left(-4m^2\pi^2(1 - r_{x*x*}(\tau))\frac{\sigma^2}{q^2}\right) \tag{16}$$

Now,

$$R_{\epsilon_q\epsilon_q}(\tau) = R_{\epsilon_q\epsilon_q}(0)r_{\epsilon_q\epsilon_q}(\tau)$$

$$= \frac{q^2}{12}\,r_{\epsilon_q\epsilon_q}(\tau), \tag{17}$$

and

$$R_{x*x*}(0) = \sigma^2.$$

($R_{xy}(\tau)$ is defined as $E\{x(t)y(t+\tau)\}$

and

$$r_{xy}(\tau) = \frac{R_{xy}(\tau)}{\sqrt{R_{xx}(0)R_{yy}(0)}} .$$

E is the expectation operator.) Substituting these expressions into equation (16) we get

$$r_{\epsilon_q\epsilon_q}(\tau) = \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{m^2} \exp\left(-4n^2\pi^2(1 - r_{x*x*}(\tau)) \frac{\sigma^2}{q^2}\right). \tag{18}$$

This equation is a monotonic decreasing function of $r_{x*x*}(\tau)$. It is shown plotted in figure 8 for σ/q = 1/3, 1/2 and 1.0. Based on this equation, the quantization error sequence can be considered a source of white noise for values of σ/q close to 1.0 and for quantizer input correlation coefficient values of 0.9 or less. Of course, the higher σ/q becomes, the higher the quantizer input correlation coefficient can be for the white noise assumption to hold.

Shaver's Results

Shaver (reference 5, pp. 7-8) derived an expression for the cross-correlation between the quantizer input and the quantization error at the same instant. Since the reference is not widely available, we will reproduce the derivation. (This will be done for zero-mean Gaussian processes and rounding.)

The cross-correlation may be written

$$R_{\epsilon_q x*}(0) = E\{\epsilon_q x*\}$$

$$= \int_{-\infty}^{\infty} \int_{-q/2}^{q/2} \epsilon_q x* p_{\epsilon_q x*}(\epsilon_q x*)\, d\epsilon_q\, dx*. \tag{19}$$

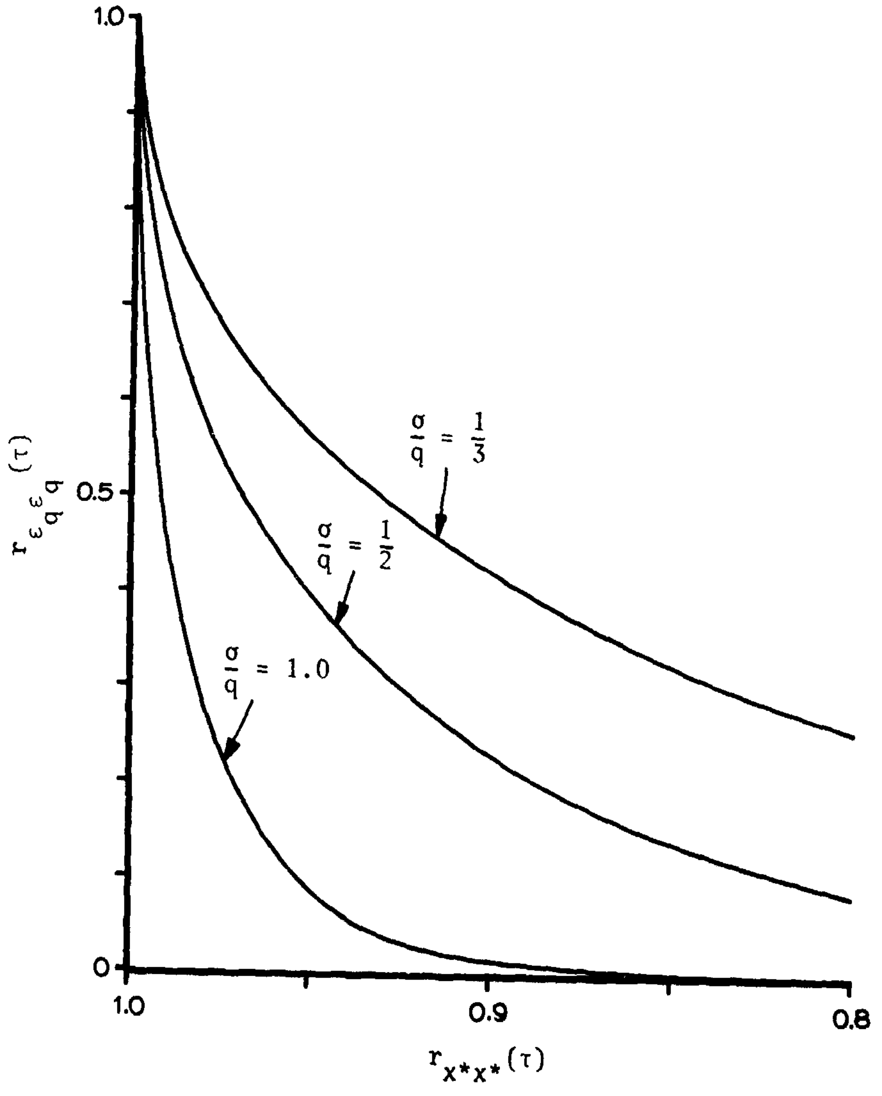

Figure 8. Correlation coefficient of quantization errors vs. correlation coefficient of successive quantizer input samples.

Substituting equation (9) into the above, we get:

$$R_{\epsilon_q x*}(0) = \int_{-\infty}^{\infty} x* \left\{ \int_{-q/2}^{q/2} \epsilon_q \sum_{i=-\infty}^{\infty} \delta(\epsilon_q - [iq - x*])\, d\epsilon_q \right\} p_{x*}(x*)\, dx*$$

$$= \int_{-\infty}^{\infty} x*\epsilon_q p_{x*}(x*)\, dx*. \tag{20}$$

The relationship between the quantization error and the quantizer input is given by equation (6). The quantization error function may be written as a Fourier series in X*,

$$\epsilon_q(x*) = \frac{q}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \sin\left(\frac{2\pi k x*}{q}\right). \tag{21}$$

Substituting equation (21) into equation (20), we get

$$R_{\epsilon_q x*}(0) = \frac{q}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \int_{-\infty}^{\infty} x* \sin\left(\frac{2\pi k x*}{q}\right) p_{x*}(x*)\, dx*. \tag{22}$$

If we consider a zero-mean Gaussian quantizer input with variance σ^2, equation (22) reduces to

$$R_{\epsilon_q x*}(0) = 2\sigma^2 \sum_{k=1}^{\infty} (-1)^k \exp\left[-2\left(\frac{\pi k \sigma}{q}\right)^2\right]. \tag{23}$$

The cross-correlation coefficient is:

$$r_{\epsilon_q x*}(0) = \frac{R_{\epsilon_q x*}(0)}{\sqrt{R_{\epsilon_q \epsilon_q}(0) R_{x*x*}(0)}}. \tag{24}$$

Since $R_{\epsilon_q \epsilon_q}(0) = q^2/12$, and $R_{x*x*}(0) = \sigma^2$, we obtain

$$r_{\epsilon_q x*}(0) = 4\sqrt{3}\,\frac{\sigma}{q} \sum_{k=1}^{\infty} (-1)^k \exp\left[-2\left(\frac{\pi k \sigma}{q}\right)^2\right]. \tag{25}$$

If $\sigma > q$, a good approximation is

$$r_{\epsilon_q x^*}(0) = -4\sqrt{3}\frac{\sigma}{q}\exp\left[-2\pi^2\frac{\sigma^2}{q^2}\right]. \tag{26}$$

This equation, like equation (15), decays extremely rapidly for $\sigma > q$. Thus, the quantization error sequence is essentially uncorrelated with the quantizer input sequence for $\sigma > q$.

MULTIPLICATION ERRORS

MULTIPLIER CHARACTERISTICS: DETERMINISTIC

The model we use for a multiplier is shown in figure 9. The multiplier input sequence x(nT), the multiplication constant J, and the multiplication output sequence y(nT) are all in the form of binary words of length N bits plus sign. Remember that y(nT) results from rounding or chopping the product x(nT) times J, which is 2N bits plus sign. The error sequence $\epsilon_c(nT)$ represents the fictitious number sequence that would be added to the product x(nT) times J in order to produce y(nT).

Table 2 shows the rounding and chopping errors for all possible combinations of x(nT) and J, and a word length of N = 2 bits plus sign. Note that the errors for J < 0 are opposite in sign from the errors for J > 0. This occurs in general for N > 2 as well. We will only consider results for J > 0 from now on. More importantly, given a particular value of x(nT), the value of the error depends on the value of the multiplication constant J. We have more to say about this after the next paragraph.

The number of values that the errors can take on depends on the word length. The number of values is:

Rounding: $2^N + 1$

Chopping: $2^{N+1} - 1$

We will use the case of table 2 as an example. We enumerate all possible errors in table 3. The number of error values are: rounding, 5; chopping, 7. Note the bounds on the errors. The rounding errors are bounded by ± one half the least significant bit (l.s.b.) in the (N + 1) bit computer word. Similarly, chopping errors are bounded by ± one l.s.b.. Previously, we showed that one l.s.b. was equivalent to the basic quantization interval q Thus, if we refer the scale of the multiplication error back to the quantizer input, the bounds on the multiplication error are the same as for the quantization error. This is why the error values are shown as an equivalent voltage referred to the quantizer input. This is the way we will show multiplication error magnitudes in the rest of this report. The value of N we use at any time is reflected in the denominator of the fractions that are used. That is, the denominator is equal to 2^N.

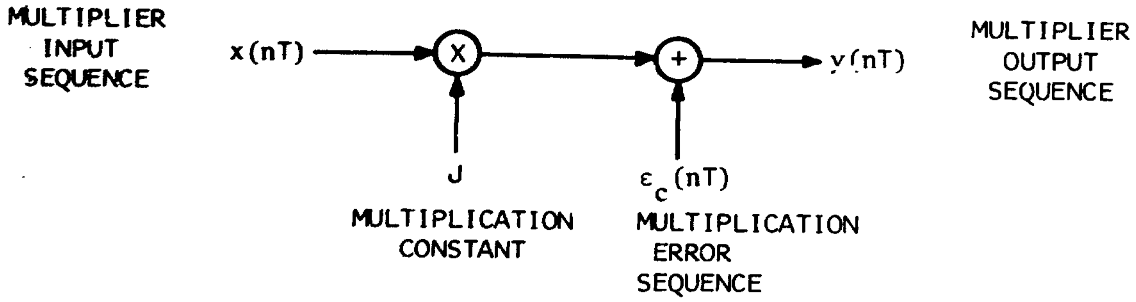
MULTIPLIER INPUT SEQUENCE
x(nT)
X
J
MULTIPLICATION CONSTANT
+
$\varepsilon_c(nT)$
MULTIPLICATION ERROR SEQUENCE
y(nT)
MULTIPLIER OUTPUT SEQUENCE

Figure 9. Multiplier model showing error generation.

Table 2. Rounding and chopping errors for all possible combinations of X and J. N = 2 bits plus sign.

EQUIVALENT VOLTAGE REFERRED TO QUANTIZER INPUT

EQUIVALENT DECIMAL FRACTION

BINARY NUMBER

x(nT) →

J ↑

			-3q	-2q	-q	0	q	2q	3q	
			-3/4	-2/4	-1/4	0	1/4	2/4	3/4	
			1.11	1.10	1.01	0.00	0.01	0.10	0.11	
3q	3/4	0.11	0.0001 0.0001	1.0010 0.0010	1.0001 0.0011	0	0.0001 1.0011	0.0010 1.0010	1.0001 1.0001	R = ROUNDING C = CHOPPING
2q	2/4	0.10	1.0010 0.0010	0	1.0010 0.0010	0	0.0010 1.0010	0	0.0010 1.0010	R C
q	1/4	0.01	1.0001 0.0011	1.0010 0.0010	0.0001 0.0001	0	1.0001 1.0001	0.0010 1.0010	0.0001 1.0011	R C
0	0	0.00	0	0	0	0	0	0	0	
-q	-1/4	1.01	0.0001 1.0011	0.0010 1.0010	1.0001 1.0001	0	0.0001 0.0001	1.0010 0.0010	1.0001 0.0011	R C
-2q	-2/4	1.10	0.0010 1.0010	0	0.0010 1.0010	0	1.0010 0.0010	0	1.0010 0.0010	R C
-3q	-3/4	1.11	1.0001 1.0001	0.0010 1.0010	0.0001 1.0011	0	1.0001 0.0011	1.0010 0.0010	0.0001 0.0001	R C

NOTE:

1.0000 = 0.0000
= 0

Table 3. Enumeration of all possible multiplication errors.
N = 2 bits plus sign.

	ERROR VALUE		
	BINARY NUMBER	EQUIVALENT VOLTAGE REFERRED TO QUANTIZER INPUT	
ROUNDING:	0.0010	$\frac{2}{4}$ q	
	0.0001	$\frac{1}{4}$ q	
	0.0000	0	$\lvert\varepsilon_c\rvert \leq \frac{q}{2}$
	1.0001	$-\frac{1}{4}$ q	
	1.0010	$-\frac{2}{4}$ q	
CHOPPING:	0.0011	$\frac{3}{4}$ q	
	0.0010	$\frac{2}{4}$ q	
	0.0001	$\frac{1}{4}$ q	
	0.0000	0	$\lvert\varepsilon_c\rvert < q$
	1.0001	$-\frac{1}{4}$ q	
	1.0010	$-\frac{2}{4}$ q	
	1.0011	$-\frac{3}{4}$ q	

POSITION OF LEAST SIGNIFICANT
BIT FOR A 2-BIT WORD LENGTH

In figure 4 we showed a quantization error function which was related to the quantizer input. It represented a mapping of the quantizer input d.d. onto the quantization error d.d. We will do the same here for multiplication errors. That is, we will construct mappings of the multiplier input d.d. onto the multiplication error d.d. Data of the form used in table 2 were used for figures 10 and 11. Figure 10 shows representative chopping error patterns for word lengths of N = 2, 3, and 4 bits plus sign. (All possible values of J are not shown). All possible values of the multiplier input are arranged along the horizontal axis. The value of the resulting multiplication error is plotted against the vertical axis. Each plot is for a particular value of J. Figure 11 is similar to 10; the main difference is that rounding error patterns are shown. In both cases, the error pattern is simply a mapping of the multiplier input value intc a corresponding multiplier error value.

We compare figures 10 and 11 with figure 4. One basic dissimilarity occurs because the multiplier input is a discrete quantity and the quantizer input continuous. If the quantizer input were discrete (in a sense, prequantized to a finer quantization interval) we would observe an error pattern similar to that for the multiplier. However, the multiplication error pattern also depends on the value of the multiplier, J. The similarity between quantization and multiplication is that the error bounds are the same: ±q/2 for rounding and ±q for chopping.

There is one special feature illustrated in figure 10 which we will use in our statistical analysis of multiplication errors. Look at the three plots starting with (N = 2, J = 1/4) on the left and ending with (N = 4, J = 4/16) on the right. The plot pattern for N = 2 is a basic pattern for longer word lengths. That is, the pattern for x = 0, 1/4, 2/4, 3/4, (N = 2) is the same pattern for x = 0, 1/8, 2/8, 3/8 (N = 3) and for x = 4/8, 5/8, 6/8, 7/8, (N = 3). Similarly, the pattern for x = 0, -1/4, -2/4, -3/4 (N = 2) is the same pattern for x = 0, -1/8, -2/8, -3/8, (N = 3) and for x = -4/8, -5/8, -6/8, -7/8 (N = 3). This same effect occurs when we go to a word length of 4 bits. The basic pattern is repeated a total of 4 times each for x > 0 and x < 0. We conclude from the figure that most error patterns are based on basic patterns. For example, suppose J = 2/16 (N = 4). We can reduce this fraction to the value 1/8 and no further. The shortest word length we can use to represent this fraction is N = 3. Thus, the basic error pattern is generated for (J = 1/8, N = 3). And the error pattern for (J = 1/16, N = 4) is a copy of this repeated according to the above procedure. The patterns in figure 11 show this same effect.

Figure 12 shows an example of a multiplier input sequence and the corresponding output and error sequences. This is similar to figure 5, the comments made for figure 5 also apply here (see p. 8). In addition, note that the result of multiplication before rounding or chopping will be less than the value of the input to the multiplier. This result depends on the value of J. The bounds on the multiplication error remain the same. This is to be contrasted with the quantizer where no operation is performed on the voltage before the quantization error is introduced.

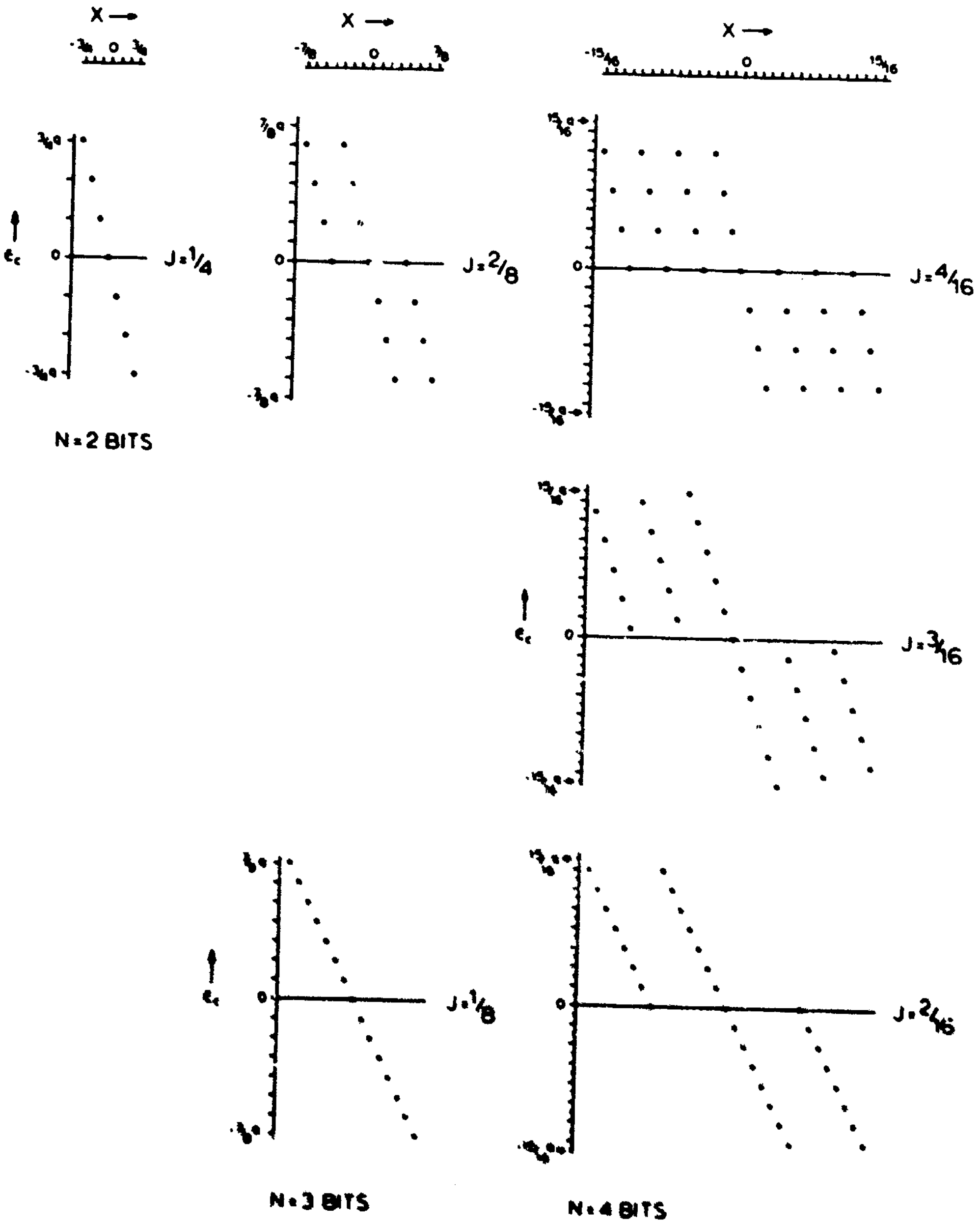

Figure 10. Chopping error patterns.

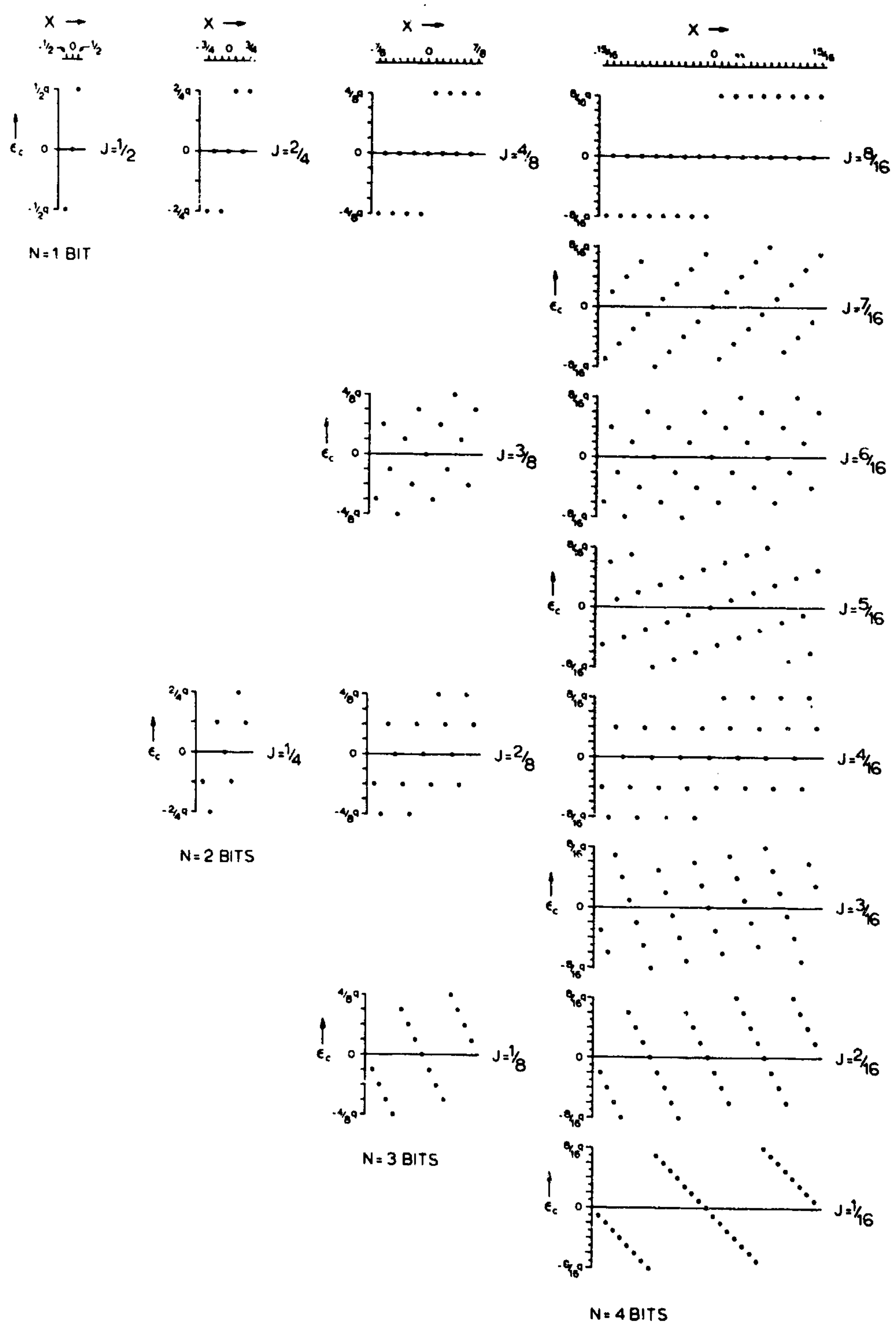

Figure 11. Rounding error patterns.

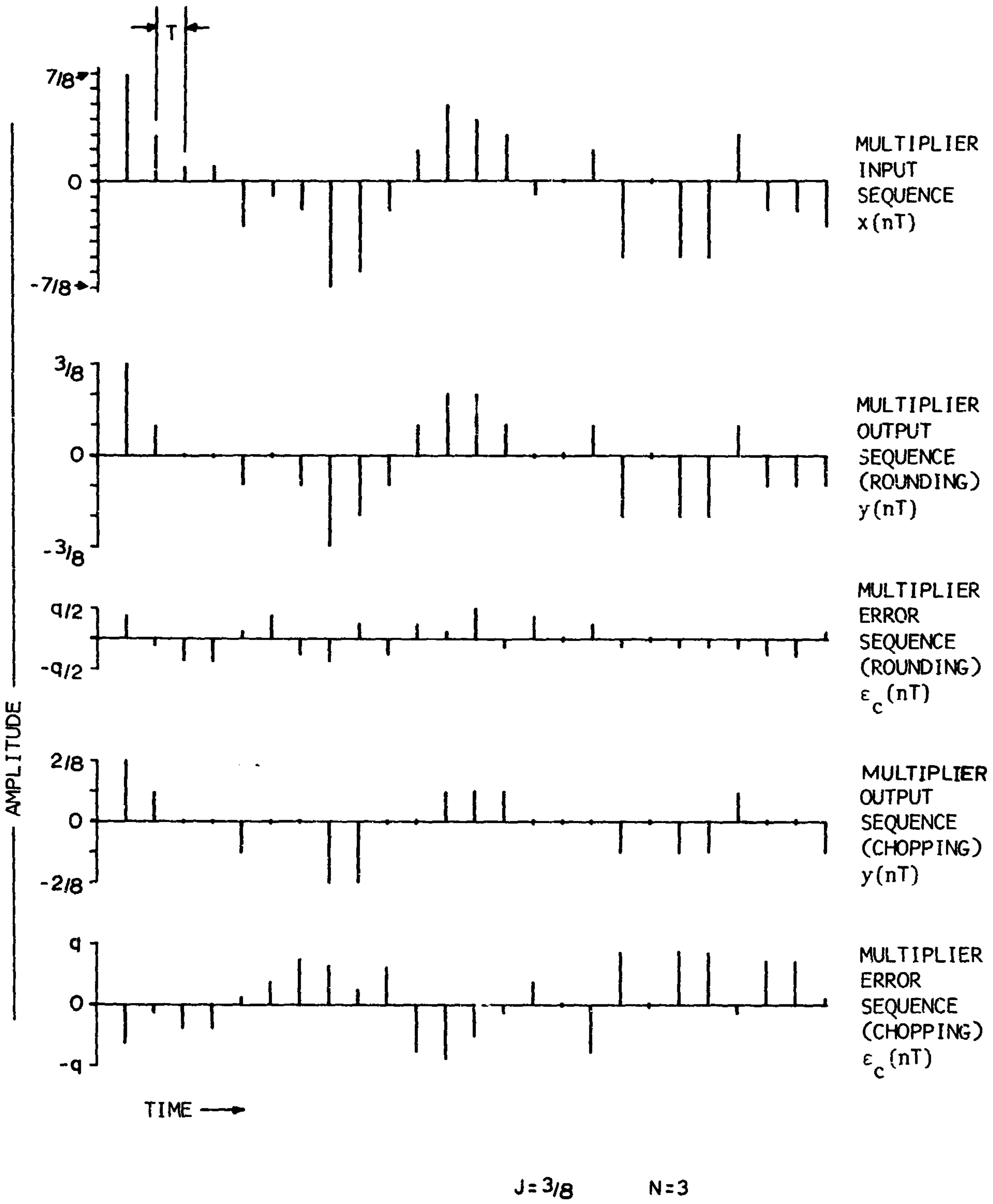

Figure 12. Example of a multiplier input sequence and multiplier output and error sequences.

DISTRIBUTION DENSITIES OF MULTIPLIER ERRORS

Figure 13 is an example of marginal and joint d.d.'s for a multiplier input and the resulting multiplication error. Chopping is shown. Figure 14 is the same except that rounding is shown. The multiplier input d.d. was derived from a zero-mean Gaussian input to a quantizer. Rounding is assumed for the quantizer. The ratio of the standard deviation to the quantization interval σ/q was 4.0. $N = 4$ and $J = 3/16$. It is clear how the multiplier input d.d. is mapped onto the multiplier error d.d. by the error pattern. The limits on the multiplication error values are obvious in both figures. Also note that the multiplication error d.d. is strongly correlated with the multiplier input d.d. We will drop consideration of chopping errors at this point.

Figure 15 shows a variety of multiplication error d.d.'s. for the same multiplier input d.d. as used for figures 13 and 14. Rounding is assumed. In some cases, the d.d.'s are approximately uniformly distributed. In other cases they are not. The dependence on the value of J is clear. These figures illustrate further the dissimilarity between multiplication errors and quantization errors.

ROUNDING ERRORS: STATISTICAL

Model For Computer Analysis

Consider equation 1. If x(nT) is a zero-mean Gaussian random process, y(nT) is also a zero-mean Gaussian random process. We will show this by an example using a first-order linear difference equation:

$$y(nT) = Ky(nT - T) - x(nT) \qquad n = 0, 1, 2, \ldots \tag{27}$$

The first value of x(nt) is x(o). We define $y(-T) = 0$. Then $y(o) = -x(o)$. Carrying out the above equation a number of times, we find that

$$y(nT) = \sum_{i=0}^{n} K^{n-i}(-x(iT)), \tag{28}$$

y(nT) consists of a linear, weighted sum of Gaussian random variables. This implies that y(nT) too is a Gaussian random variable. Also, since

$$E[x(nT)] = 0, \qquad E[y(nT)] = 0.$$

We can use this same procedure for other forms of equation 1. All of this assumes that quantization and computation errors are not present. We have shown that quantization of a Gaussian random variable results in a discrete d.d. which is approximately

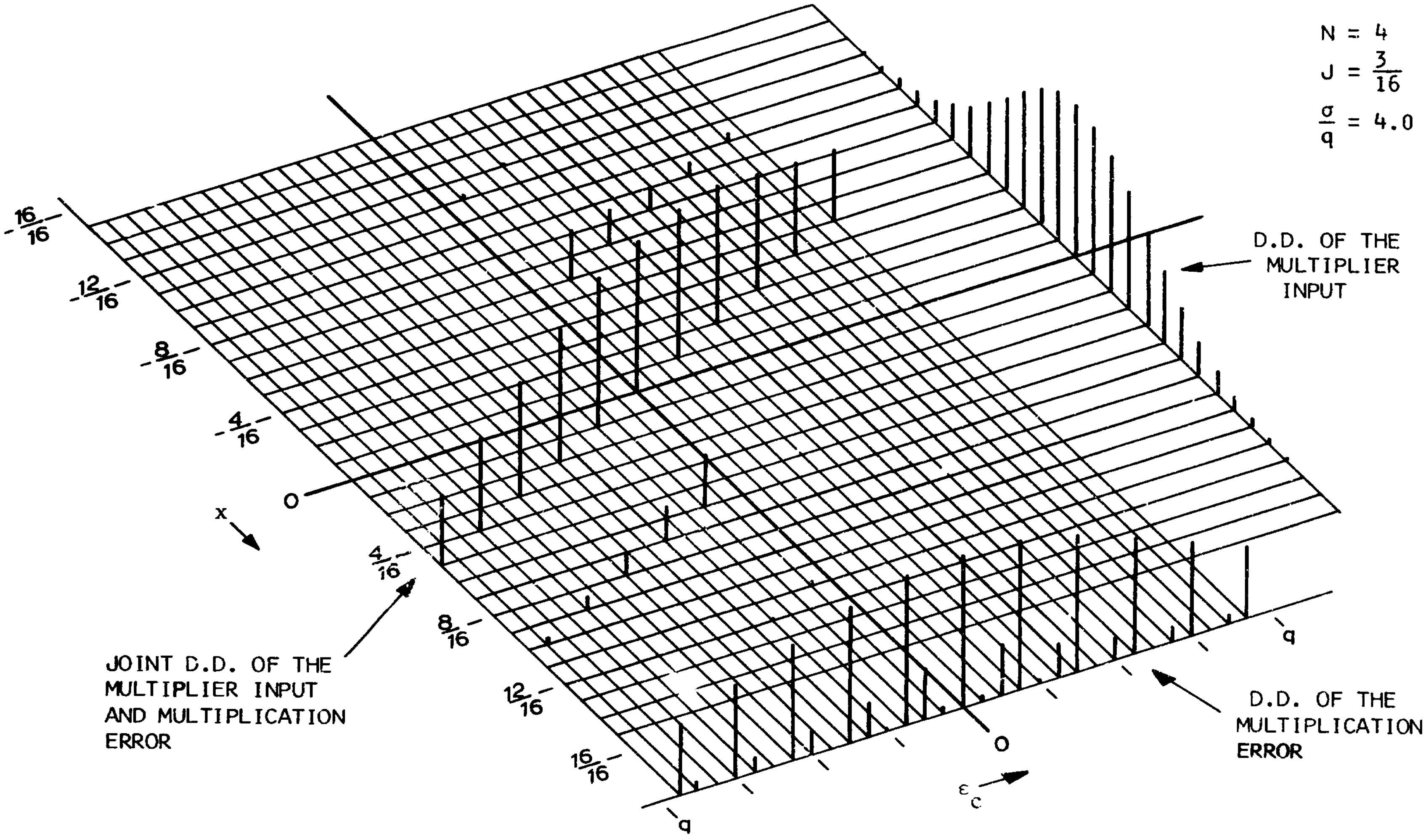

Figure 13. Example of marginal and joint d.d.'s; chopping.

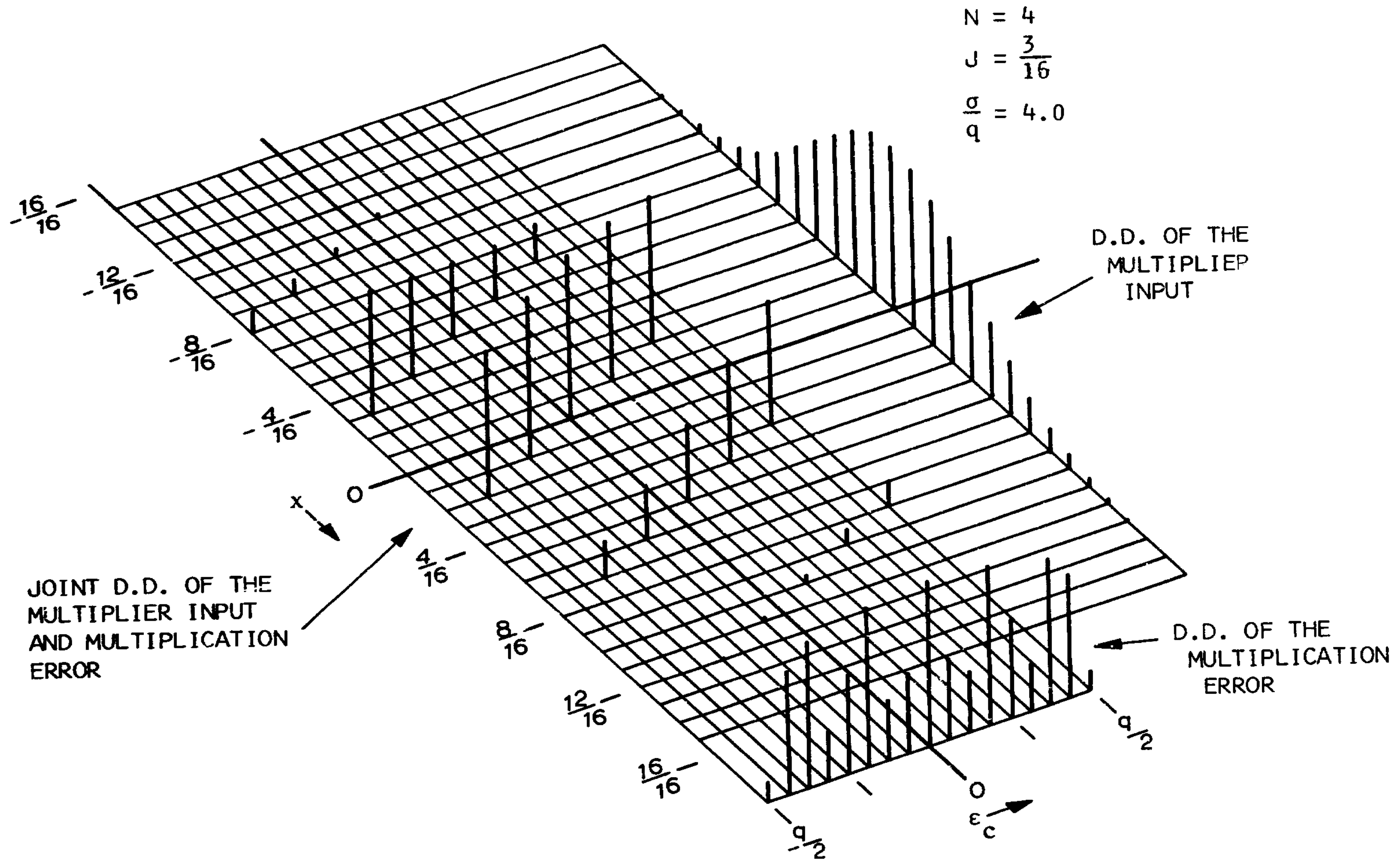

Figure 14. Example of marginal and joint d.d.'s; rounding.

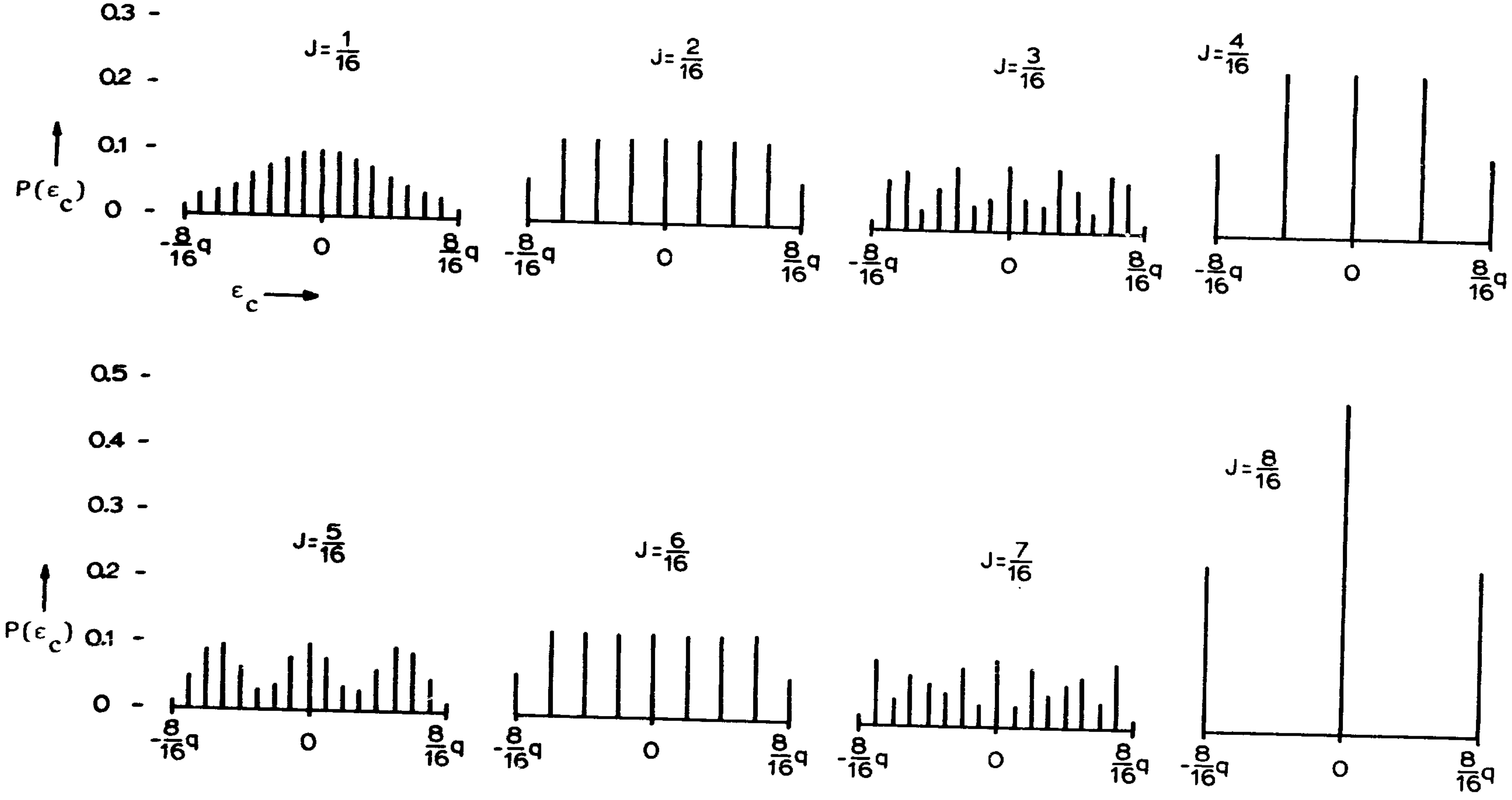

Figure 15. Examples of multiplication error d.d.'s. N - 4, σ/q = 4.0, and rounding is assumed.

Gaussian in appearance. We will assume this same d.d. for the input and output of a digital filter. The output d.d. will only be approximate due to the addition of computational errors. Thus, in equation (1), the x(nT) and y(nT) are assumed to be governed by the same d.d. That is, they are the result of quantizing (with rounding) a Gaussian sequence of random variables. With these assumptions we simplify our analysis model to that shown in figure 9.

The analyses described below have been programmed and run on a CDC 1604 computer.

Multiplication Error Variance

The variance of the computation error was computed as

$$\mathrm{Var}(\epsilon_c) = E[(\epsilon_c - \bar{\epsilon}_c)^2]$$

$$= \sum_{i=-2^N+1}^{2^N-1} \epsilon_{c_i}^2 \, P_{x'}(iq). \qquad (29)$$

The variance depends on the following parameters: J, N, and σ/q. ϵ_{c_i} is the computation error that results when the multiplier input is equal to $i2^{-N}$. (Remember that the decimal form of the ADCON output is $i2^{-N}$ when the quantizer input is in the interval $(iq \pm q/2)$.) As we have shown, the value of ϵ_{c_i} depends on the value of J.

The variance was computed for the following parameters: $N = 7$; $J = i2^{-N}$ $(i = 1, 2, ..., 2^7 - 1)$; $\sigma/q = 2.0, 4.0, 8.0, 16.0$. Results are shown in figure 16. The variance for a uniformly distributed error is shown as a horizontal line in the center of the graphs. The interval of ±10 percent of this value is also shown.

The first graph is for $\sigma/q = 2.0$. Except for a small range of values for $J \sim 32/128$, 64/128, and 96/128 most of the values of error variance fall outside the 10 percent interval. For comparison, the proportional error in the variance of the *quantizing* error for $\sigma/q = 1.0$ is about one part in 10^7. So, we see that the performance of the computation error variance is very much worse than that of the quantizing error variance. As σ/q is increased, the computation error variance converges to the value of a uniformly istributed error for most values of J. For certain values of J the error variance converges to some other value. (See values for J = 16/128, 32/188, ..., 112/128.)

The data points in figure 16 are shown connected for ease in visualization. This does not imply that we are safe in interpolating variance values for N > 7 and values of J

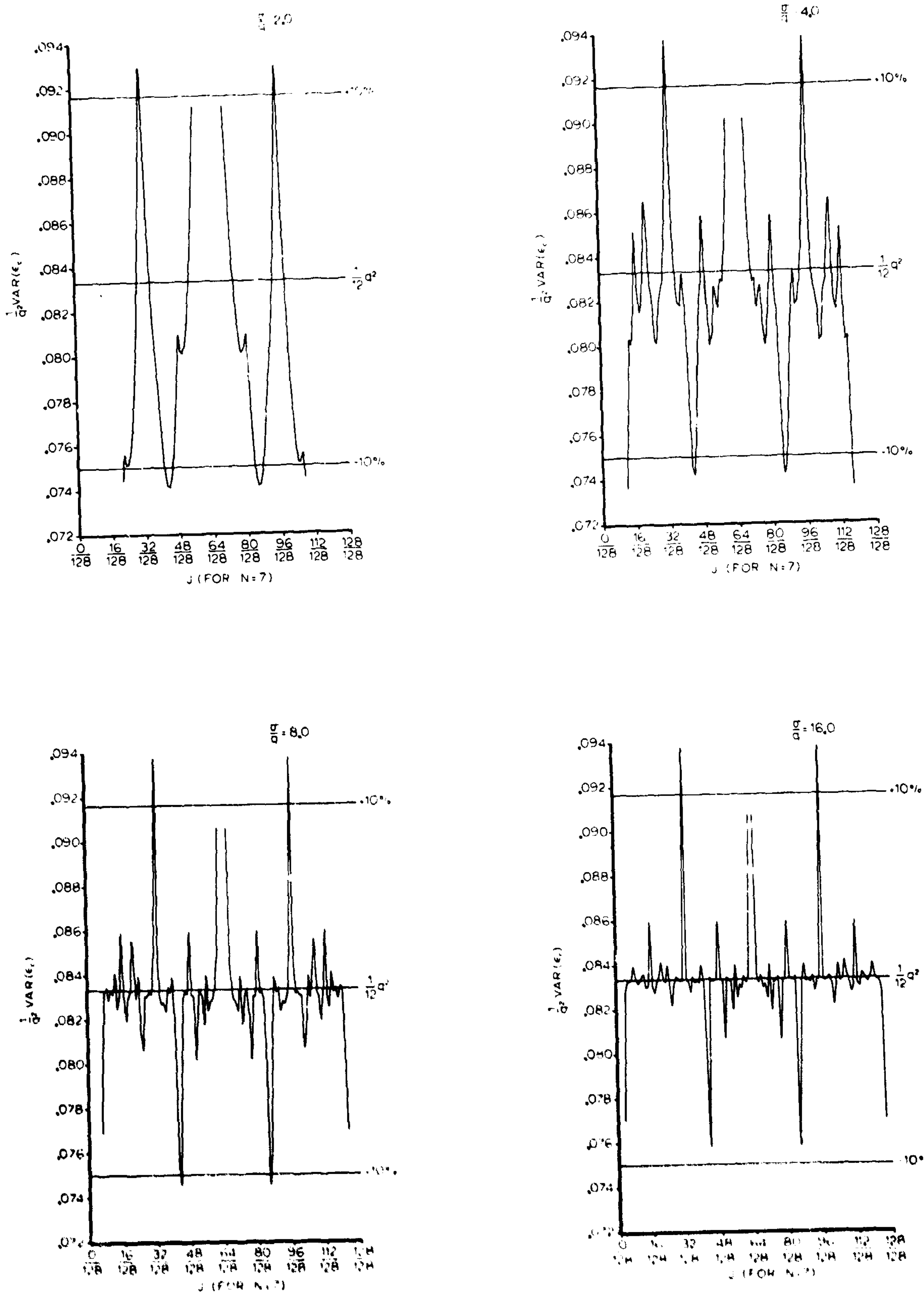

Figure 16. Multiplication error variance as a function of the multiplication constant J (N = 7).

that are representable only for N > 7. For example, we could have plotted variances for N = 6. Then, the two downward spikes at J = 43/128 and J = 85/128 would not have been indicated.

There is an upper limit to the standard deviation allowable for the input to a multiplier. This limit is set by the size of N, and is imposed at two points. The first is the input to the quantizer. If the quantizer input exceeds $(2^N - 1/2)q$ volts overloading occurs. The second point is the output of the digital filter represented by equation (1). It is possible for the sum of all the terms to be too big for the word size (≥1.0 in this case). (That is, overflow occurs.) In the practical case, we can limit the standard deviation so that quantizer overloading or register overflow will occur infrequently. A good value for the standard deviation is one-fourth the voltage which is equivalent to the maximum size of the word. The probability of overloading in either the positive or negative sense is then about 6.33×10^{-5}. In figure 16, the maximum value of σ/q that can be accommodated for N = 4 is $\sigma/q = 4.0$. Other maximum values are N = 5, $\sigma/q = 8.0$; N = 6, $\sigma/q = 16.0$; N = 7, $\sigma/q = 32.0$; etc. We call these limits on σ/q the 4σ load limits.

Autocorrelation Computations

The autocorrelation coefficient for two successive computation errors is

$$r_{\epsilon_c\epsilon_c}(T) = \frac{E\left\{\epsilon_c(n_1T)\epsilon_c(n_2T)\right\}}{E\left\{\epsilon_c^2(n_1T)\right\}}$$

$$= \frac{\sum\limits_{i=-2^N+1}^{2^N-1}\sum\limits_{j=-2^N+1}^{2^N-1} \epsilon_{c_i}\epsilon_{c_j} P_{x'x'}(iq,jq)}{\sum\limits_{i=-2^N+1}^{2^N-1} \epsilon_{c_i}^2 P_{x'}(iq)}, \qquad 1 = |n_1 - n_2| \qquad (30)$$

where

$$P_{x'x'}(iq,jq) = \text{Prob}\left\{x'(n_1T) = iq,\ x'(n_2T) = jq\right\}$$

$$= \text{Prob}\left\{(i - \tfrac{1}{2})q < x^*(n_1T) \leq (i + \tfrac{1}{2})q,\ (j - \tfrac{1}{2})q < x^*(n_2T) \leq (j + \tfrac{1}{2})q\right\}$$

$$= \int_{(i-1/2)q}^{(i+1/2)q} \int_{(j-1/2)q}^{(j+1/2)q} P_{x^*x^*}(x^*(n_1T), x^*(n_2T))\,dx^*(n_1T)\,dx^*(n_2T). \qquad (31)$$

and $P_{x^*x^*}(x^*(n_1T), x^*(n_2T))$ is a bivariate Gaussian d.d. with autocorrelation $r_{x^*x^*}(T)$. $P_{x'x'}(iq, jq)$ was obtained numerically using Simpson's rule. (Note that the results described are for a given autocorrelation value of the quantizer input, *not* the multiplier input.)

The autocorrelation coefficient was computed for the following parameters: $N = 6$; $J = i2^{-N}$ $(i = 1, 2, ..., 2^6 - 1)$; $\sigma/q = 1, 2, ..., 6$; $r_{x^*x^*}(\tau) = 0.9$. $r_{\epsilon_c\epsilon_c}(\tau)$ is plotted as a function of J in figure 17. Here too, the data points are shown connected for visual effect only. (Results for $\sigma/q > 6.0$ were not obtained due to the amount of computer time needed.

The values are also much higher than the value of the quantizing error correlation coefficient for $\sigma/q = 1.0$ (see figure 8; when $r_{x^*x^*} = 0.9$, $r_{\epsilon_q\epsilon_q} = .0117$). In figure 17, for $\sigma/q = 6.0$, only two values of J (13/64 and 51/64) yield a lower value. Other values of J yield lower values for $\sigma/q < 6.0$. But the computation error correlation coefficient later comes back up when $\sigma/q = 6.0$. The coefficient values seem to stabilize for $J = 16/64$, 32/64 and 48/64 for even this restricted range of σ/q. Unfortunately, since we do not have data for $\sigma/q > 6.0$, we can only speculate that the correlation coefficient will be low enough for most cases of interest. It is not low enough, for the most part, when $\sigma/q \leqslant 4.0$, nor for $N \leqslant 4$ when the 4σ load limits are taken into account.

Cross-Correlation Coefficient

The cross-correlation coefficient between the multiplier input x(nT) and the resulting multiplier error $\epsilon_c(nT)$ at the same instant is

$$r_{\epsilon_c x}(0) = r_{\epsilon_c x'}(0)$$

$$= \frac{E\{x'(nT)\epsilon_c(nT)\}}{\sqrt{E\{[x'(nT)]^2\}\ E\{[\epsilon_c(nT)]^2\}}}$$

$$= \frac{\sum_{i=-2^N+1}^{2^N-1} iq\epsilon_{c_i}P_{x'}(iq)}{\sqrt{\sum_{i=-2^N+1}^{2^N-1} (iq)^2P_{x'}(iq)\ \sum_{j=-2^N+1}^{2^N-1} \epsilon_{c_j}^2P_{x'}(jq)}}. \tag{32}$$

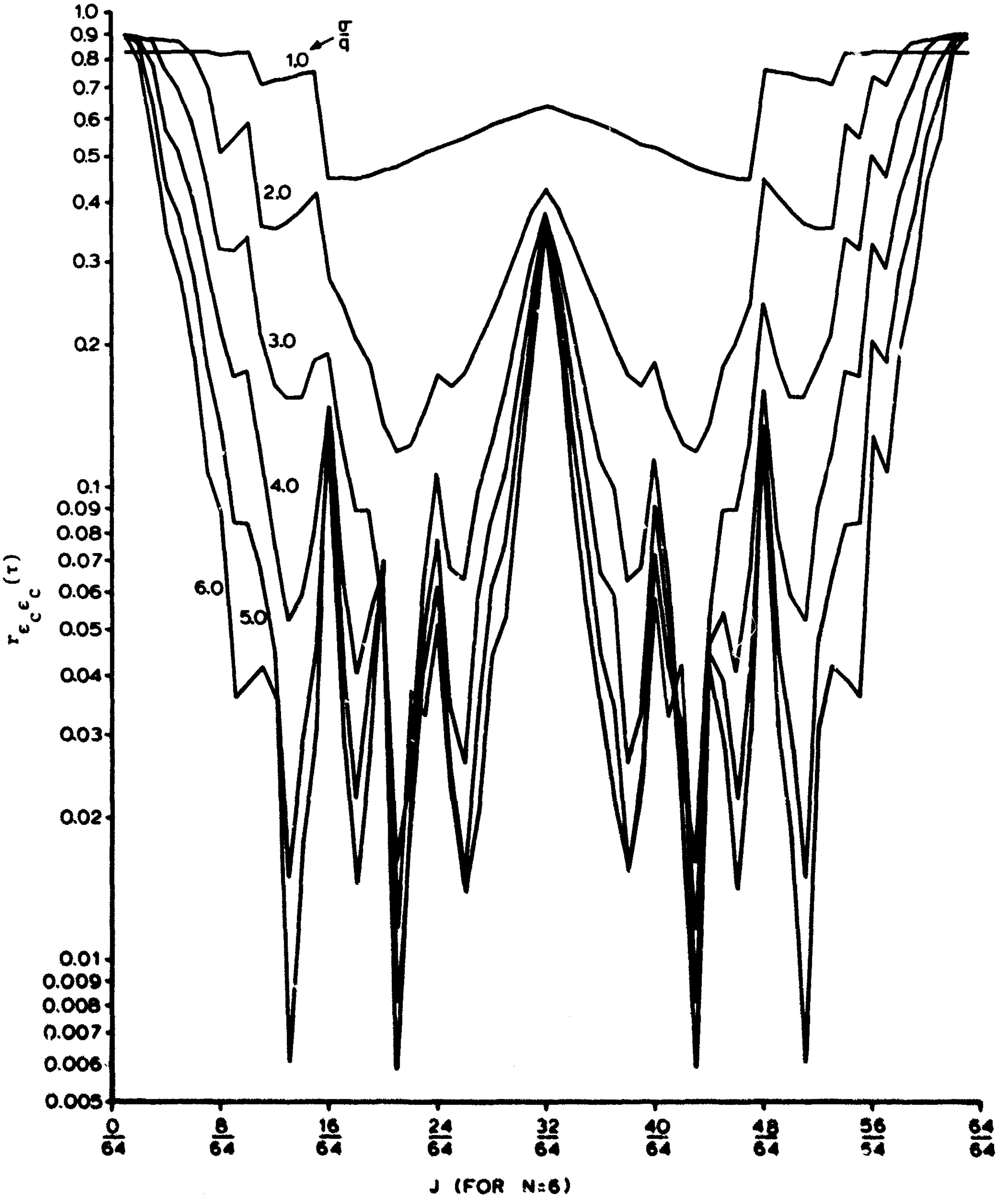

Figure 17. Autocorrelation coefficient of multiplication errors vs. J and σ/q. Value of autocorrelation coefficient of quantizer input is 0.9.

The cross-correlation coefficient was computed for the following parameters: $N = 6$; $J = i2^{-N}$ $(i = 1, 2, \ldots, 2^6 - 1)$; $\sigma/q = 1, 2, \ldots, 30$. Its behavior as a function of σ/q is evident in figure 18.

Each curve tends to vary wildly starting with $\sigma/q = 1.0$. Then, it settles down to some non-zero positive value. The reason for the positive value can be traced to the appearance of the error patterns in figure 11. When $x > 0$, the mapping onto ϵ_c occurs for more positive values of ϵ_c than negative values. When $x < 0$, the mapping occurs for more negative values of ϵ_c than for positive values. Each curve settles more quickly for those values of J that are representable for $N < 6$. $J = 32/64$ is the most extreme example. Next comes $J = 16/64$ and $48/64$. Then, $J = 8/64$, $24/64$, $40/64$ and $56/64$, and so on. Values of J representable only for $N \geq 6$ seem to have the least tendency to settle down in the range shown for σ/q. This behavior is correlated with the behavior of the multiplication error variance as a function of σ/q.

The rate of settling down seems to be inversely correlated with the final non-zero value of the cross-correlation coefficient. That is, the faster it settles down, the farther from zero it stays as σ/q becomes large.

Generalization of these results for $N > 6$ is not always safe. Figure 19 shows why. It is a plot of the cross-correlation coefficient as a function of J for $\sigma/q = 30.0$. The plot is in two parts. The top part is for values of J that are representable for both $N = 6$ and 7. The bottom part is for only those values of J that are representable for $N = 7$. The top is uniform in appearance. But, it is obvious that pitfalls occur if we try to extrapolate performance for $N > 6$ for numbers that are only representable for $N > 6$. The bottom part shows additional variations that are not predictable by looking at the top part. They are also significantly non-zero in some cases when we consider that $\sigma/q = 30.0$ is very close to the 4σ load limit for $n = 7$. Incidentally, the spikes that occur for $J = 43/128$ and $J = 85/128$ are at the same position as the downward spikes in figure 16 for $\sigma/q = 16.0$.

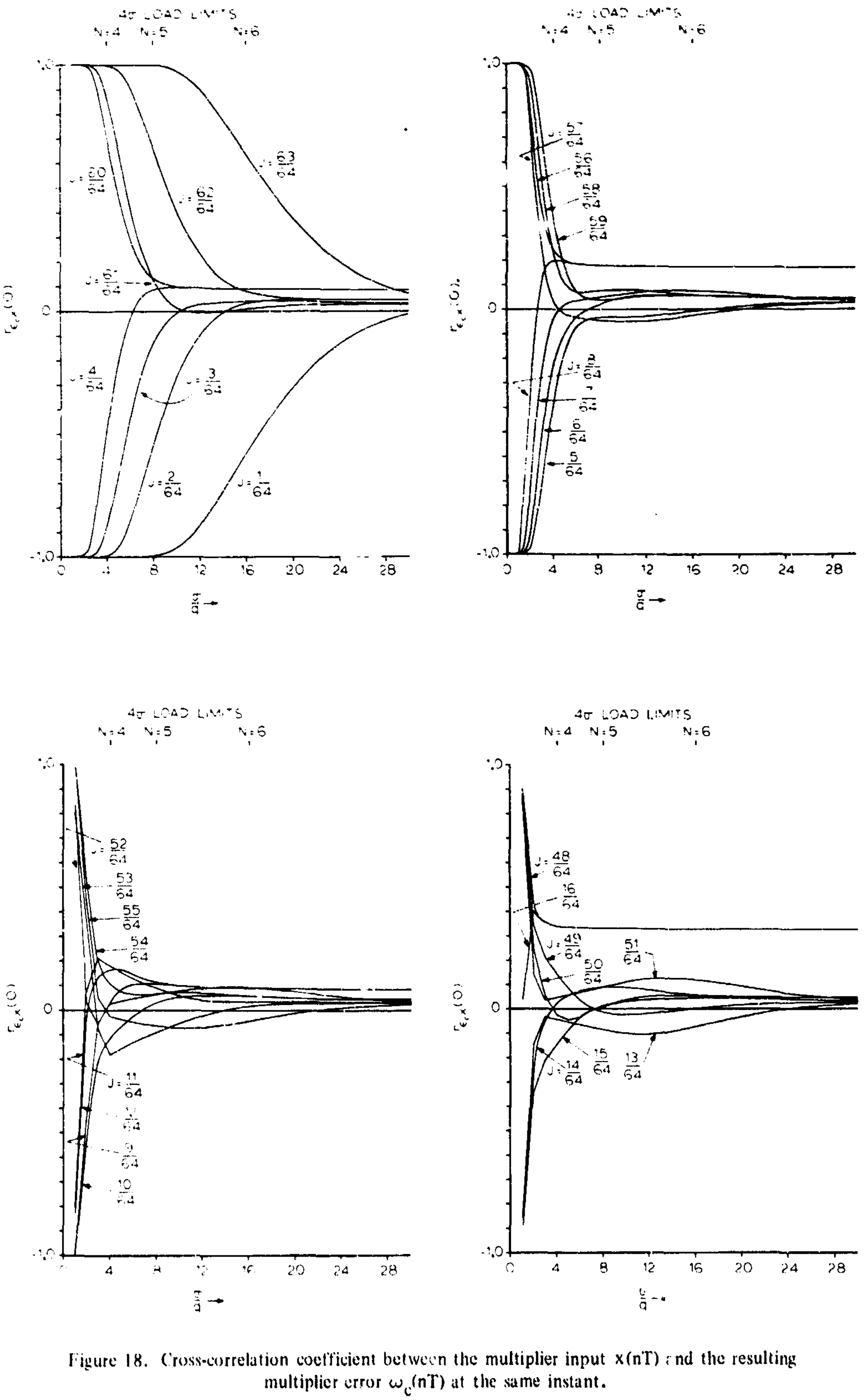

Figure 18. Cross-correlation coefficient between the multiplier input x(nT) and the resulting multiplier error ω_c(nT) at the same instant.

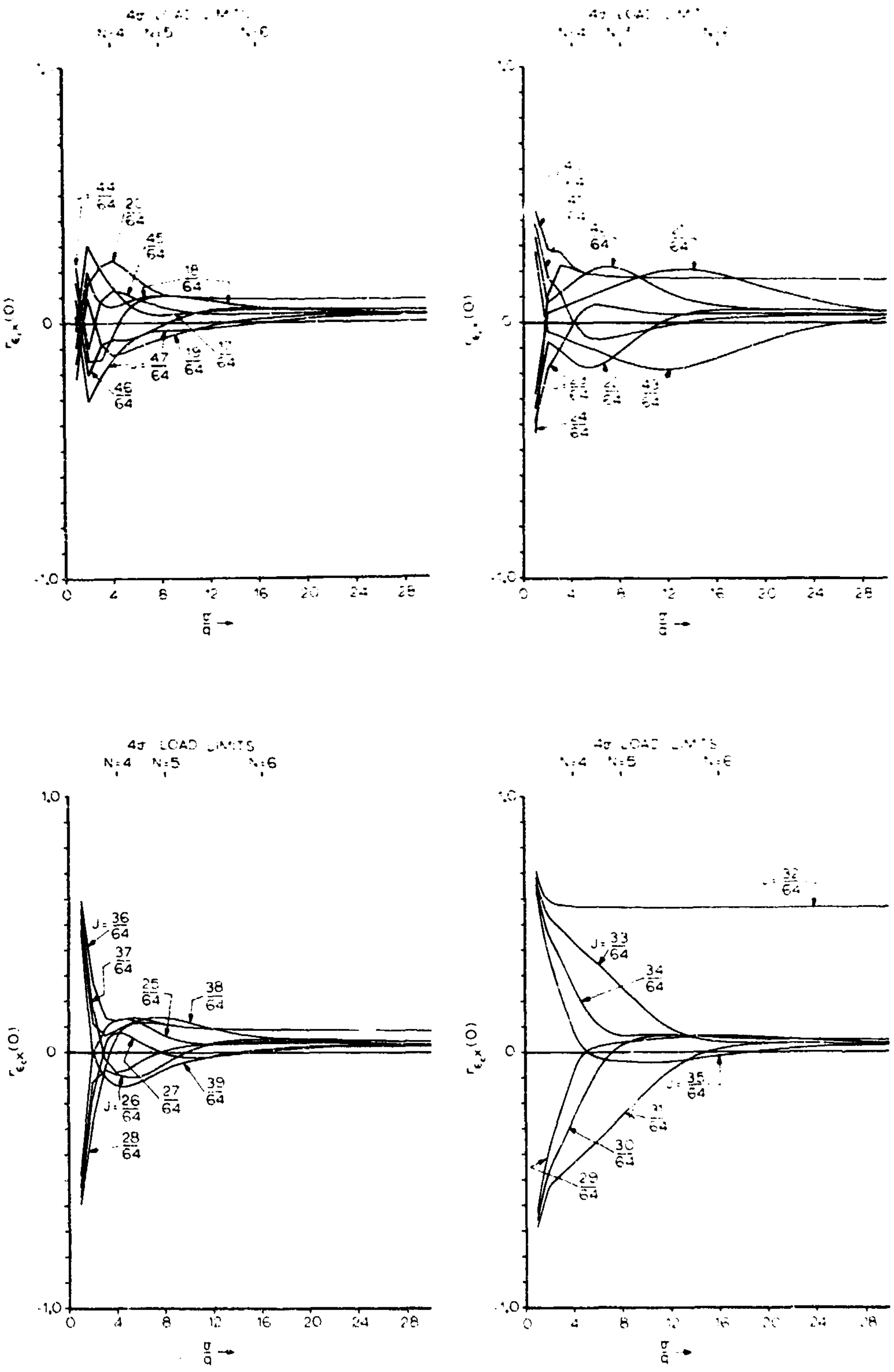

Figure 18. Continued.

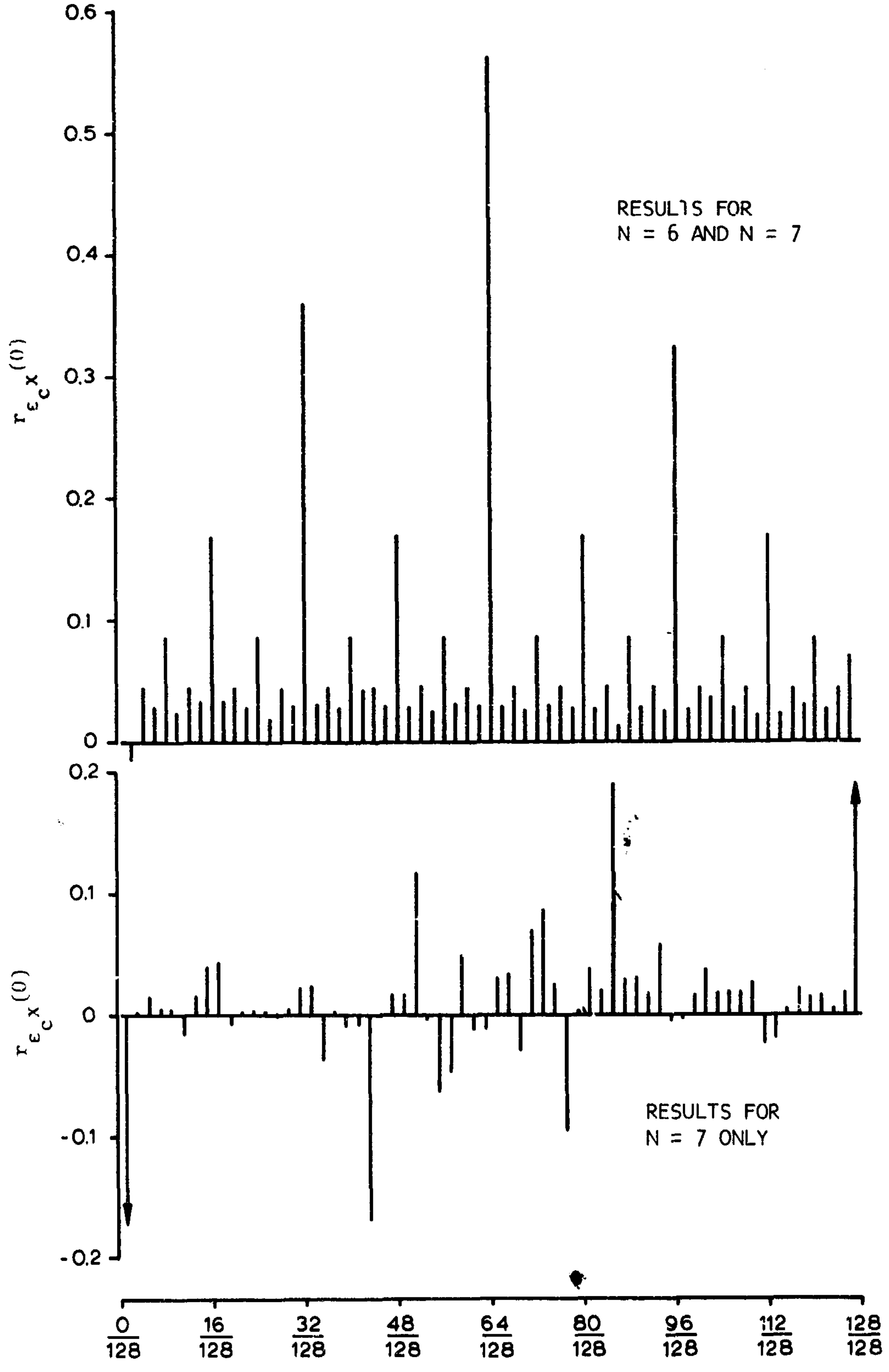

Figure 19. Cross-correlation coefficient between multiplier input and multiplication error vs. J, σ/q = 30.0 .

SUMMARY

We have examined the following statistical properties of quantization and multiplication rounding errors:

1. d.d. of the error,
2. variance of the error,
3. autocorrelation coefficient between successive error values, and
4. cross-correlation coefficient between the quantizer or multiplier input with the resulting errors.

Specific results were obtained for zero-mean Gaussian random processes as follows.

QUANTIZATION ERRORS

The statistics of quantization errors depend only on the ratio of the process standard deviation to the quantization interval size (σ/q). The mapping of the quantizer input d.d. onto the quantization error d.d. is continuous. Consequently, for $\sigma/q \geqslant 1.0$, the error d.d. is almost exactly uniform between $\pm q/2$, and the error variance is very near $q^2/12$. Both the autocorrelation and the cross-correlation coefficients were negligible. Furthermore, the equations show that the quantization error approaches arbitrarily close to $q^2/12$ as σ/q increases, while the autocorrelation and cross-correlation coefficients approach arbitrarily close to zero.

MULTIPLICATION ROUNDING ERRORS

For rounding errors, the above properties depend not only on σ/q, but on the word size, N, and the value of the multiplier, J, as well. Furthermore, the discrete nature of the computer word causes a discrete mapping of the multiplier input d.d. onto the multiplication error d.d. Consequently, for the limited range of parameters considered, most values of J yield an error d.d. which is not uniform in the continuous sense but shows a variance approaching $q^2/12$. Similarly, most autocorrelation and cross-correlation values approach zero, but stabilize at some non-zero value as σ/q becomes large. However, some values of J result in large, non-zero autocorrelation and cross-correlation values and a variance which diverges widely from $q^2/12$.

Shortcomings of Present Analysis

It should be remembered that the results are based on data obtained for digital words of size $N \leqslant 7$ bits plus sign bit. We have not considered the possibility of rounding off fewer bits leaving a result of multiplication which is greater in size than the multiplier input word size. Neither have we obtained data for word sizes greater than 8 bits. Operation with word sizes greater than 8 bits is of interest due to the increased availability of process control computers in the 12- and 16-bit word-size range. (11 bits plus sign and 15 bits plus sign.)

An Approach to Further Analysis

Since generalization of the results of this paper to word sizes greater than 8 bits has its shortcomings, we suggest the following approach. First, assume ideal multiplication error statistics (shape of d.d., variance size, and auto- and cross-correlation coefficients) for the analysis of the effect of multiplication errors. Analysis approaches are worked out in reference 2. Once the preliminary design is fixed, perform a Monte Carlo simulation of the digital filter in a digital computer. Then compute the error d.d., the error variance, and auto- and cross-correlation coefficients for each multiplier coefficient in the filter. Do this for a representative set of filter input sequences. (Sequences of correlated Gaussian random variables are easily generated using computer programs.) If the results for a multiplier coefficient are bad, it may be possible to get good results by using a slightly different coefficient value. Of course, the change in filter characteristics would have to be acceptable.

REFERENCES

1. G. Maley and E. Skiko. *Modern Digital Computers,* Prentice Hall, 1964.

2. B. Gold and C. Rader. *Digital Processing of Signals,* McGraw-Hill Book Co., Inc., 1969.

3. W. R. Bennett. Spectra of Quantized Signals, *Bell System Technical Journal,* v. 27, pp. 446-472, July 1948.

4. B. Widrow. A Study of Rough Amplitude Quantization by Means of a Nyquist Sampling Theory. *Institute of Radio Engineers Transactions on Circuit Theory,* v. CT-3, no. 4, pp. 266-276, December 1956.

5. Stanford Electronics Laboratories. Technical Report 7050-5, Topics in Statistical Quantization, by H. N. Shaver. May 1965.

6. J. B. Knowles and R. Edwards. Finite Word Length Effects in Multirate Direct Digital Control Systems. *Proceedings of the Institute of Electrical Engineers* (London), v. 112, pp. 2376-2384, December 1965.

7. J. B. Knowles and R. Edwards. Complex Cascade Programming and Associated Computational Errors, *Electronics Letters,* v. 1, no. 6, pp. 160-161, August 1965.

8. J. B. Knowles and R. Edwards. Effect of a Finite-Word-Length Computer in a Sampled-Data Feedback System, *Proceedings of the Institute of Electrical Engineers* (London), v. 112, no. 6, pp. 1197-1207, June 1965.

9. B. Gold and C. M. Rader. Effects of Quantization Noise in Digital Filters, American Federation of Information Processing Societies, *Proceedings, Spring Joint Computer Conference,* pp. 213-219, 1966.

10. Massachusetts Institute of Technology Research Laboratory of Electronics. Technical Report 465, Analysis of Digital and Analog Formant Synthesizers, by B. Gold and L. R. Rabiner. 28 June 1968.

11. A. Papoulis. *The Fourier Integral and its Applications*, McGraw-Hill Book Co., Inc., 1962.

12. B. Widrow. Statistical Analysis of Amplitude Quantized Sampled-Data Systems, *American Institute of Electrical Engineering Transactions on Application and Industry*, No. 52, pp. 555-568. January 1961.

LIST OF SYMBOLS

$A,-B$: upper and lower limits (volts) of the sampler and quantizer

α: a multiplicative relation between $x'(nT)$ and $x(nT)$

$\delta(x)$: the delta function

$\epsilon_c(nT)$: multiplication error sequence

$\epsilon_{c_i}(nT)$: the multiplication error that results when the multiplier input is equal to $i2^{-N}$

$\epsilon_q(nT)$: quantization error sequence

ϵ_ν: error of the variance of the quantization error

ϵ'_ν: proportional error of the variance of the quantization error

f_s: sampling frequency

K_i, L_i: digital filter coefficients.

N: the number of bits in a digital word (excluding the sign bit)

ϕ: "frequency"

$\phi(t)$: any continuous function

q: the size of the basic quantization interval (volts)

$R_{xy}(\tau)$: the cross-correlation between any two random variables x and y; the autocorrelation of x when y = x

$r_{xy}(\tau)$: the cross-correlation coefficient between any two random variables x and y; the autocorrelation coefficient of x when y = x

σ^2:	the variance of a (Gaussian) d.d.
T:	the time interval between analog waveform samples and numbers of a number sequence
$\tilde{x}(t)$:	analog waveform
$x^*(nT)$:	analog waveform sampler output sequence
$x(nT)$:	ADCON output sequence: the input sequence to a digital filter
$x'(nT)$:	quantizer output sequence
$y(nT)$:	digital filter output sequence

SPECIAL TERMS

ADCON:	analog-to-digital converter
d.d.:	(probability) distribution density
l.s.b.:	least significant bit